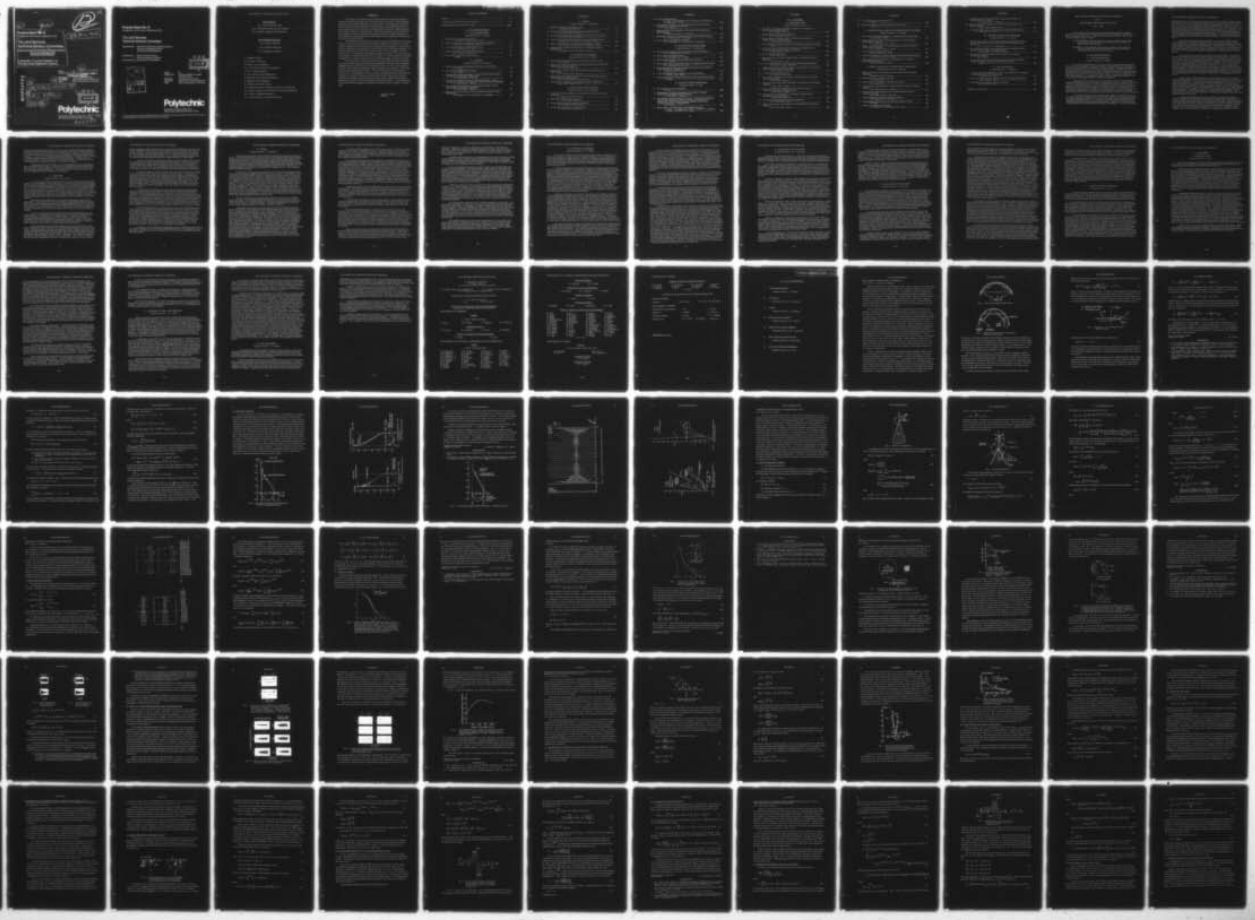
AD-A034 764 POLYTECHNIC INST OF NEW YORK BROOKLYN MICROWAVE RESE--ETC F/G 20/3
PROGRESS REPORT NUMBER 41, 15 SEPTEMBER 1975 THROUGH 14 SEPTEMB--ETC(U)
NOV 76 A A OLINER F44620-74-C-0056
UNCLASSIFIED POLY-MRI-452.41-76 NL

1 OF 7
AD
A034764



FG

12

(6) Number

Progress Report No. 41

15 September 1975 through 14 September 1976

to

The Joint Services Technical Advisory Committee.

Representing: ~~The Air Force Office of Scientific Research~~
~~The U.S. Army Research Office~~
~~The Office of Naval Research~~

(9) Interim rept.,

A summary of current research at
The Microwave Research Institute.

(15)

Report R-452.41-76

for Contract F44620-74-C-0056

(16) ~~Project No.~~ 4751

(11) Submitted November 1976

This document has been approved for public release and sale; its distribution is unlimited.

(14) POLY-MRI-452.41-76

(10) Arthur A. Oliner

(12) 64 p.

ADA034764

DDC
RECEIVED
JAN 24 1977
D

Polytechnic

Polytechnic Institute of New York
333 Jay Street, Brooklyn, New York 11201

4473 LB
409138

Progress Report No. 41

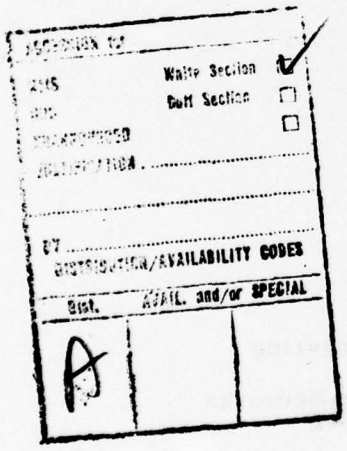
15 September 1975 through 14 September 1976

The Joint Services Technical Advisory Committee

Representing: The Air Force Office of Scientific Research
The U.S. Army Research Office
The Office of Naval Research

Submitted by: Arthur A. Oliner, Director
Microwave Research Institute
Coordinated by: Jerome Fox and Jean Maher

DDC
RECEIVED
JAN 24 1977
D



Report R-452.41-76

Submitted November 1976

for Contract F44620-74-C-0056
Project No. 4751

This document has been approved for public release and sale; its distribution is unlimited.

Polytechnic

Polytechnic Institute of New York
333 Jay Street, Brooklyn, New York 11201

Qualified requestors may obtain additional copies from the Defense Documentation Center; all others should apply to he Clearinghouse for Federal Scientific and Technical Information

MICROWAVE RESEARCH INSTITUTE

DIRECTORATE

A. A. Oliner: Director, MRI

J. Fox: Operations and Government Relations

S. W. Rosenthal: Budgets and Laboratories

MRI RESEARCH COUNCIL

A. A. Oliner, Chairman

J. Fox, Executive Secretary

S. W. Rosenthal, Ex Officio

L. Bergstein: Optics

H. L. Bertoni: Acoustics

A. Hessel: Electromagnetics

H. J. Jurerschke: Solid State and Materials

L. Kurz: Communications

E. Levi: Electric Power Engineering

N. Marcuvitz: Wave-Matter Interactions

A. Papoulis: Data Processing

B. Senitzky: Solid State Devices

M. L. Shooman: Safety, Reliability and Software Engineering

E. J. Smith: Computers and Computer-Communication Networks

W. T. Walter: Quantum Electronics

D. C. Youla: Systems, Control and Networks

PREFACE

This report to the Joint Services Technical Advisory Committee presents a summary of the research programs in the broad field of electronics conducted during the past year at the Polytechnic Institute of New York. These programs are pursued within the framework of the Microwave Research Institute, and they involve the academic research activities of faculty in the departments of Electrical Engineering and Electrophysics, Physics, and Chemistry. The research projects cover a broad spectrum ranging from basic theoretical investigations in physics, applied mathematics, and engineering to experimental efforts involving basic measurements and the development of devices and materials.

The format of this annual report permits a coherent presentation of the various phases of the Joint Services Electronics Program (JSEP) at the Polytechnic and their relation to ongoing research in electronics sponsored by other agencies. This presentation is intended for the information of the Air Force Office of Scientific Research, the Army Research Office and the Office of Naval Research and, in addition, the other sponsors who are individually acknowledged throughout the report. The principal aims of the JSEP are to initiate deserving lines of research in a timely fashion and to develop investigations to a stature sufficient to attract individual support on their own merits.

In the early days of the Microwave Research Institute, the research program consisted primarily of projects involving electromagnetics and microwave components. Although the name of the Institute has remained the same, the nature of the research programs has broadened substantially, and the programs now encompass a wide range of topics within the field of electronics. The current programs are organized into twelve areas: electromagnetics; acoustics; optics; quantum electronics; solid state and materials; wave-matter interactions; electric power engineering; communications; computers and computer-communication networks; safety, reliability and software engineering; systems, control and networks; and data processing. A short description of the nature of these programs is presented in the Introduction, on pages xi through xxxii.

Arthur A. Oliner
Director

PRECEDING PAGE BLANK-NOT FILMED

TABLE OF CONTENTS

Preface iii

The Microwave Research Institute Programs xi

Participating Faculty and Research Staff xxxiii

I. ELECTROPHYSICS

A. ELECTROMAGNETICS

Program Director: A. Hessel

1. Excitation of Large Concave Surfaces
L. B. Felsen, A. Green and A. Hessel 2

2. Radiation by Sources in a Unidirectionally Conducting Plane
A. Hessel, S. Siddiqi and J. Shmoys 6

3. Transient Diffraction by a Semi-Infinite Cone 15
K. K. Chan and L. B. Felsen.....

4. Analysis of Surface Waveguide Discontinuities 22
E-W. Hu and L. Bergstein.....

5. Stray Field of a Slotted Microstrip Line 27
I. Palócz.....

B. ACOUSTICS

Program Director: H. L. Bertoni

1. Physical Process of SAW Schottky Diode Memory Correlator
W-C. Wang 30

2. Experiments on SAW Diode Correlator
W-C. Wang 34

3. Rayleigh Wave Fields Excited by an Acoustic Beam Incident from a Liquid Near the Rayleigh Critical Angle 41
Y. L. Hou and H. L. Bertoni .....

4. Rayleigh Wave Scattering from a Liquid-Filled Crack: Part I - Variational Expressions for Reflection and Transmission Coefficients
Y. L. Hou and H. L. Bertoni 50

5. Rayleigh Wave Scattering from a Liquid-Filled Crack: Part II- Edge Conditions and Numerical Results
Y. L. Hou and H. L. Bertoni 57

6. Stability and Noise Analysis of the Acoustoelectric Phase-Locked Loop Device
H. Schachter, F. A. Cassara and L. Rosenheck..... 66

CONTENTS

C. OPTICS

Program Director: L. Bergstein

1. Specular Reflection from Sodium Vapor
   V. Nangia and B. Senitzky 76
2. Waveguides for Integrated Optics Formed by Metal Platings
   A. A. Oliner and S. T. Peng 81
3. Ray Optical Calculation of Edge Diffraction in Unstable Resonators
   L. B. Felsen and C. Santana 91
4. Analysis and Experimental Evaluation of Lossy Grating Couplers
   S. Austin, S. T. Peng, F. T. Stone and T. Tamir 102
5. Effects of Imperfections on the Efficiency of Optical Couplers
   D. W. Fradin, P. K. Cheo, S. T. Peng and T. Tamir 106
6. RF-Sputtering of Aluminum-Oxide Films for Integrated Optics Applications
   L. Bergstein, E-W. Hu and M. Eschwei 109

D. QUANTUM ELECTRONICS

Program Director: W. T. Walter

1. Composition and Equilibrium Total Pressure of Bismuth Vapor
   W. T. Walter 116
2. Double-Pulse Discharge Examination of Bismuth Vapor for Efficient, Pulsed Laser Action at 4722Å
   W. T. Walter and K. Park 120
3. Spontaneous-Emission Transition Probability of the 4722A Atomic Bismuth Line
   W. T. Walter and N. Solimene 126
4. Application of the Heisenberg Picture to the Averaged Description of the Propagation of Optical Beams
   M. C. Newstein and D. Ramakrishnan 131
5. Two-Photon Correlations in a Wide Angle Chaotic Light Beam
   S. M. Turner and D. B. Scarl 136

E. SOLID STATE AND MATERIALS

Program Director: H. J. Juretschke

1. Avalanche Initiated Oscillations in Unipolar Structures under Pulsed Excitation
   B. Senitzky and S. Gottfried 139
2. Hall Effect in Silicon Inversion Layers
   Y. L. Yao and B. Senitzky 146
3. Large Signal Transient Response of MOSFETS
   R. Kinasewitz and B. Senitzky 152
4. Reliable Ohmic Contacts to Silicon Devices
   M. Eschwei and S. Gottfried 155

CONTENTS

5. New Solid State Materials
E. Banks, M. Greenblatt, B.R. McGarvey, S. Nakajima, M. Shone and G. Torre 160

6. A Simple New Technique for the Measurement of the Elastic and Magnetoelastic Properties of Ferrites
L.M. Silber 163

7. Magnetoconductivity Modulation of Permalloy Film by Substrate Electrostriction
P. Mazumdar and H.J. Juretschke 166

8. Modulation of Thin Film Conductivity by Substrate Electrostriction
D. Lischner and H.J. Juretschke 170

9. On the Origin of the Metallic Field Effect
H.J. Juretschke, D. Lischner and P. Mazumdar 173

F. WAVE-MATTER INTERACTIONS

Program Director: N. Marcuvitz

1. The Interaction of Intense Laser Radiation with Metal Surfaces
M.C. Newstein and N. Solimene 177

2. Self-Focusing of Coherent Pulses
M.C. Newstein and F. Mattar 181

3. Analysis of Stimulated Brillouin Scattering in Plasmas
E.S. Cassedy 184

4. Observation of Wave Packet Bifurcation in a Magneto-Plasma Column
E.E. Kunhardt and B.R. Cheo 190

5. Resonant Coupling of Ionization Waves and Acoustic Gravity Waves in the Presence of a Magnetic Field
H. Eun and S.H. Gross 198

6. A Kelvin-Helmholtz Type Instability in Plasma Striations
H. Kudyan and K. Chung 213

7. Thermal Analysis of Vacuum Vessel Bake-out of Fusion Test Reactor
K. Chung 216

G. ELECTRIC POWER ENGINEERING

Program Director: E. Levi

1. New Concepts in Linear Propulsion
E. Levi 220

2. Magnetostatic Potential and Magnetic Flux Density Along the Armature of Homopolar Inductor Machines
E. Levi and M.S. Gemolos 226

3. New Results in Electrical Machine Theory: Formulas of the Blasius Type for the Calculation of Force and Torque on the Rotor of a Cylindrically Uniform Electromagnetic Machine Viewed as an Aerofoil Immersed in a Magnetic Fluid
D.C. Youla 230

4. Static Tests on the Claw Pole Linear Synchronous Motor
E. Levi, L. Birenbaum, P. Chan, D. Fontana, F. Lalezari, J. Mullen, N. Rokkos and D. Wong 234

CONTENTS

II. SYSTEMS

A. COMMUNICATIONS

Program Director: L. Kurz

1. The Class of Lehmann Alternatives as Means for Evaluation of Performance in Robust Detection
   B. Gotz and L. Kurz 242
2. Signal Design and Rank M-ary Detection
   Y. C. Ching and L. Kurz 249
3. On Improvement in Performance of Transmission Systems in Additive Gaussian and Burst Noise
   L. Kurz and R. J. Nawrocky 260
4. On Estimation of Channel Conditions in Digital Communication Systems
   L. Kurz 268
5. A Robust Vector Robbins-Monro Procedure with Application to Parameter Estimation in m-Interval Detectors
   P. Kersten and L. Kurz 275
6. M-level Quantile Detectors
   B. Gotz and L. Kurz 285
7. Signal Selection and Optimum Signal Design Using Robust Partition Statistics
   P. Kersten and L. Kurz 292
8. A Novel FM Detector for Suppression of Interchannel Interference
   F. A. Cassara and H. Schachter 300
9. Optimization of M-ary Digital Transmission Systems in the Presence of Intersymbol Interference and Noise
   L. Kurz and R. J. Nawrocky 305
10. A Generalized Quantile Detector
    J. I. Cochrane and L. Kurz 315
11. The Nyquist Problem in the Presence of Nonlinearities in the Data Transmission System
    L. Kurz and G. Soloway 322
12. Edge Detection Using ANOVA Techniques in Conjunction with Quantile Statistics
    L. Kurz and P. Legakis 328
13. Gray Level Detection Using M-Interval Occupancy Vector in Conjunction with Mann-Whitney Statistics
    L. Kurz and P. Legakis 337
14. Deterioration in Performance of a Duobinary PAM System by Introduction of Nonlinearities
    L. Kurz and M. Wernicki 348
15. Bivariate Dependent Sequential Partition Detectors
    R. F. Dwyer and L. Kurz 353
16. Relative Performance of Sequential and Fixed Sample Size Partition Detectors
    R. F. Dwyer and L. Kurz 358

CONTENTS

17. A Flexible Model for Impulsive Noise
L. Kurz 365

18. Suboptimum Detection in Nonstationary Gaussian Noise
L. Kurz 371

B. COMPUTERS AND COMPUTER-COMMUNICATION NETWORKS

Program Director: E. J. Smith

1. On the Partitioning Representation of Finite Sets and Its Application
T. Lee and E. J. Smith 374

2. A Semi-Automatic Design Verification System for Communications Line Control Procedures
E. J. Lancevich and E. J. Smith 385

3. CPDL -- A Communications Protocol Design Language: Application to Full-Duplex Protocols
E. J. Lancevich and E. J. Smith 391

4. On a Technique for Dynamic Routing
R. R. Boorstyn and A. Livne 397

5. Channel Assignments for Cellular Mobile Telecommunications Systems
R. R. Boorstyn and R. J. Pennotti 399

6. Queueing Analysis of Internodal Acknowledgement in Computer-Communication Networks
R. R. Boorstyn and J. Stark 409

C. SAFETY, RELIABILITY AND SOFTWARE ENGINEERING

Program Director: M. L. Shooman

1. Effects of Manpower Deployment and Error Generation on Software Reliability
M. Shooman and S. Natarajan 426

2. Statistical Theory of Program Testing
A. E. Laemmel 436

3. Statistical Theory of Program Testing and Proving
A. E. Laemmel 445

4. Program Paths and the Number of Tests Needed to Verify a Computer Program
G. S. Popkin and M. L. Shooman 458

5. Application of Logical Operations to Finding Feasible Paths in a Program Flowchart
G. S. Popkin and M. L. Shooman 466

6. Proposed Measures for the Evaluation of Software
S. N. Mohanty and M. Adamowicz 469

7. Seeding/Tagging Estimates of the Number of Software Errors
B. Rudner 476

8. Modular Programming Techniques
E. Lipshitz, M. L. Shooman and H. Ruston 484

CONTENTS

9. Implementation of Shooman's Model of Exhaustive Testing: An Automatic Type 1. A Tester
D. L. Baggi and M. L. Shooman 491

10. Optimal Inspection Schedules for Failure Detection When Tests Hasten Failures
L. Shaw and N. Wattanapanom. 497

11. Reliability Applications of a Multivariate Exponential Distribution
L. Shaw and C-L. Hsu 502

D. SYSTEMS, CONTROL AND NETWORKS

Program Director: D. C. Youla

1. New Results in System Theory: Bauer-type Factorization of Positive Matrices and the Theory of Matrix Polynomials Orthogonal on the Unit Circle
D. C. Youla and N. N. Kazanjian..... 507

2. On the Design of Single-Loop Single-Input-Output Feedback Control Systems in the Complex-Frequency Domain
J. J. Bongiorno, Jr. and D. C. Youla 534

3. Generalized Image Restoration by the Method of Alternating Orthogonal Projections
D. C. Youla 544

4. An Approach to the Pattern Classification and Estimation Problem Using the Concept of Stochastic Differential Equations
L. Kurz 560

E. DATA PROCESSING

Program Director: A. Papoulis

1. A Robustized Kiefer-Wolfowitz Procedure with Application to Threshold Estimation for Block Data Processing
P. Kersten and L. Kurz..... 571

2. Recursive Factor Analysis Methods in Feature Extraction Problems
L. Kurz and C. S. Yoon 579

3. Nonparametric Probability Density Estimation
I. M. Habib and L. Kurz 586

Publications and Reports 593

THE MICROWAVE RESEARCH INSTITUTE PROGRAMS

of the

POLYTECHNIC INSTITUTE OF NEW YORK

(1975 - 1976)

This introductory section contains an over-all summary of the research programs in the broad field of electronics conducted at the Polytechnic Institute of New York during the past year. These programs are organized below under the following descriptive subject headings:

ELECTROPHYSICS: Electromagnetics; Acoustics; Optics; Quantum Electronics; Solid State and Materials; Wave-Matter Interactions; and Electric Power Engineering.

SYSTEMS: Communications; Computers and Computer-Communication Networks; Safety, Reliability and Software Engineering; Systems, Control and Networks; and Data Processing.

I. ELECTROPHYSICS

A. ELECTROMAGNETICS

Program Director: A. Hessel

The various investigations which fall under the broad heading of electromagnetics involve the propagation, guiding, radiation, and diffraction of electromagnetic waves in a large variety of environments. Included under this category during the past few years are major programs in wave types near interfaces, radiation from and scattering from periodic structures, various antenna investigations, particularly those involving planar and conformal phased arrays, radiation from sources and scattering by obstacles in media with relatively arbitrary properties, a systematic exploitation of ray-optical techniques in propagation, guiding, scattering and antenna problems, and topics related to biological hazards at microwave frequencies.

The program on wave types near interfaces some years ago introduced the concept of leaky waves, showed its value in the explanation of many radiation phenomena, and laid the foundations for a new class of traveling-wave antennas, the leaky-wave antenna, which permitted better agreement between theoretical and measured radiation patterns than any other type of antenna. The program led to a very general study of wave types, including several categories of complex guided wave and lateral wave, and showed their interrelations. The lateral wave was also shown to be the mechanism which permits most of the point-to-point communication in a jungle or forest environment.

A current investigation which relates to wave types near interfaces involves the lateral beam shift encountered by a beam of finite width when incident upon an interface under appropriate conditions. If the beam is incident upon the interface between two dielectric half-spaces at or very near to the critical angle of total reflection, the resulting beam shift is called the Goos-Hänchen shift. This shift is a special case of a larger class which is currently under systematic investigation. At the critical angle

of total reflection, the incident beam couples to a lateral wave which propagates along the interface. If a layer is present on the interface, or if a periodic grating is located there, and if the beam is incident at the angle corresponding to a leaky wave on the layer or grating, a very pronounced lateral beam shift is found. This beam shift, which is related to a pole in the wavenumber spectral plane, is much larger in general than the beam shift due to the lateral wave, which corresponds only to a branch point. This beam-shift mechanism has recently been used to achieve high coupling efficiencies at optical frequencies between laser beams and thin films. The beam shift itself offers a simple condition for optimizing the coupling parameters. The general properties of such beam shifts and their application to optical coupling devices are under systematic study.

The lateral beam shift is also required for a correct ray-optical model for modal propagation in thin films and optical fibers. Such modal propagation studies, which also involve the excitation of guiding structures of this type by Gaussian beams, and the optical coupling studies mentioned above, are considered in more detail in the Optics section. The leaky-wave beam shift is also shown to permit the optimization of the coupling efficiency of acoustic wedge transducers and couplers from acoustic surface waves to linear acoustic waveguides; these studies are reported in the Acoustics section.

The program on periodic structures has many ramifications. One phase relates to general studies of radiation from surfaces or interfaces which are modulated periodically, and in this context the Brillouin diagram for radiating periodic structures was first introduced and its usefulness clarified. This phase also included studies of various types of periodically-modulated slow-wave antennas. Another phase involved the study of special symmetries in periodic structures, particularly screw and glide symmetry. Studies of scattering by periodic surfaces led to a new and physically-satisfying theory of Wood's anomalies on optical gratings, an effect which remained poorly understood for 60 years. This theory shed substantial light on scattering resonances in general, and these ideas have been subsequently applied to other areas, such as scattering from plasmas and radiation from phased-array antennas. These studies are now being extended to the investigation of effects due to the boundedness of finite beams.

Another study relating to periodic structures led to a new way of treating mutual-coupling effects in phased arrays which takes these effects into account rigorously and automatically, and it proposed the first compensation scheme for the minimization of such effects. A recent study extended this technique to the explanation of resonance effects in large phased arrays, a phenomenon that can cause the array to become unexpectedly "blind" at certain scan angles. This theoretical and experimental study achieved a rather complete understanding of this phenomenon--its causes and how it may be avoided. In addition, a guided-wave approach has been applied specifically to the analysis of a slot-fed phased array on a conducting circular cylinder. This analysis is the first one for such a structure in which mutual-coupling effects have been completely taken into account.

A concentrated effort is under way to establish methods of analysis for mutually-coupled arrays of aperture elements on non-planar conducting convex and concave surfaces, using both modal and asymptotic (ray) methods. Non-planar arrays of this type are called "conformal arrays" because they are often located on the skins of aircraft, where they must "conform" to the shape of the structure (wing, fuselage, etc.) on which they are located. The motivation for this effort is twofold: a) for application to a recent interesting development in antenna arrays called the "Dome Antenna", which is a feed-array combination of a single planar phased array with a

passive, dome-shaped, feed-through lens, and b) for the design of flush-mounted slot arrays on conical or ogive surfaces for homing missile applications.

While the design of conformal arrays generally requires the knowledge of the coupling coefficients and element patterns in a mutually-coupled convex environment, a conformal feed-through scanning lens array design necessitates, in addition, the knowledge of mutual coupling for arrays on concave conducting surfaces. The following topics are currently being pursued: (1) the analysis of dually-polarized open-ended circular waveguide elements, and (2) the coupling coefficients in cylindrical arrays of waveguide-fed axial slits.

Analysis shows that the decay of the coupling coefficients in arrays on concave surfaces may differ radically from that on planar arrays. The decay rate may be considerably slower (even for $ka \sim 100$, where a is the local equivalent cylindrical radius). This feature indicates that the use of a planar array matching network may not be adequate, and that the size of the test array used for the measurement of coupling coefficients on concave surfaces should be larger than that used in the planar case.

The design of homing missile slot arrays requires the knowledge of coupling coefficients and element patterns on conical or ogive conducting surfaces. A computer program based on the harmonic series solution for a conical geometry is extremely time consuming. To alleviate this difficulty, analytical expressions based on principles of the Geometric Theory of Diffraction (GTD) have been obtained which include modifications to torsional geometry of the existing GTD formulas valid for the torsionless case. Based on these expressions, a computer program has been developed for the calculation of the GTD radial or circumferential slot far-field patterns on conical surfaces. The numerical results compare well with those based on the harmonic series, except in the forward axial (tip) region, and in regions of low intensity in the GTD expressions, where the tip scattering that was neglected in the GTD analysis is of importance. For the computation of mutual coupling, GTD expressions were derived for the surface current due to a magnetic current element located on a conical surface, based on the asymptotic solution of the canonical problem of a circular cylinder. For axial slots, the mutual admittance values calculated via the GTD expressions for a circular cylindrical surface for $ka \sim 10$ check in the deep shadow and for the near neighbors with harmonic series results found in the literature. For circumferential slots, the asymptotic expressions must include additional truncation functions.

The current program includes the following topics: the analysis of spiral elements for phased array applications; the performance of a two-dimensional scanned lens array; modal analyses of mutual coupling on concave spherical surfaces; ray methods for two- and three-dimensional concave surfaces; and the synthesis of wide-angle scanning lens arrays.

The program on propagation and scattering in media with arbitrary properties was motivated by the recognition that recent technology requires a knowledge of the propagation of electromagnetic waves in physical environments of increased complexity, and of the radiation characteristics of antennas and the scattering characteristics of obstacles in these environments. Examples of such media are provided by ionized plasmas, either in the laboratory or in outer space, whose electrical properties can be described macroscopically in terms of an inhomogeneous or homogeneous isotropic dielectric, an anisotropic dielectric, or a mechanically deformable dielectric material, the choice of a particular model being dependent on the circumstances in question. In addition, turbulent processes in the medium may require the inclusion of statistical properties in its description. When a plasma medium surrounds an antenna or a scattering object, a situation encountered when a rocket or satellite passes through the

ionosphere or when a high speed vehicle re-enters the upper atmosphere, the above-mentioned processes of radiation and scattering become relevant for problems of radio communication and detection. They are also relevant for optical detection of satellites and other objects since the optical signal traverses an inhomogeneous and turbulent atmosphere on its way to the ground-based detector.

In these studies, ray-optical concepts play an important role. The exploration of the range of validity of ray-optical procedures in propagation, guiding, radiation and scattering in a general environment, and the subsequent application of these techniques to problems of current interest, provide the basis for much of this quasi-optic phase of the electromagnetics program.

Particular attention in the quasi-optic program has been given to the following subject areas: 1) propagation of pulses in a dispersive environment; 2) scattering by discontinuities in homogeneously and inhomogeneously filled waveguides or ducts; 3) propagation and scattering of exponentially-decaying (evanescent) fields in lossless or lossy regions by complex ray techniques; 4) inhomogeneous wave tracking. To deal with the first category, an asymptotic theory based on space-time rays (a generalization of geometric-optical rays) and plane-wave dispersion surfaces has been developed. Complicated transient wave processes are thereby described in terms of the motion, evolution, and interaction of wave packets. The theory has been applied to spatially-varying, temporally-varying and moving plasma media, and the results obtained are relevant for the study of pulse degradation, pulse compression and related phenomena.

In the second category, novel applications of ray-optical (i.e., high-frequency) methods have been found to yield remarkably accurate results for radiation from, and scattering by, discontinuity elements in guiding structures, even in the relatively-low-frequency propagation range of only the dominant mode. The technique has recently been applied to VLF mode excitation and conversion in the earth-ionosphere waveguide, and has now been extended to multiwave media such as compressible plasmas or elastic solids. Moreover, by viewing stable and unstable optical resonators as open-ended waveguides along the coordinate transverse to the mirror axis, a novel and promising analytical method for determining losses and field configurations in open resonators is presently under study. The method has so far been remarkably successful in explaining, by a simple model, the complicated loss behavior of eigenmodes in unstable resonators, which are widely employed with high-power lasers. This study is discussed further in the Optics section.

Emphasis in the third category is placed on determining the local propagation properties of evanescent fields for the purpose of developing therefrom a theory of propagation and scattering analogous to the geometrical theory of diffraction for non-evanescent fields. While evanescent fields can be regarded as traveling along ray trajectories in complex space, the investigation seeks to clarify the physical significance of "complex rays" and their possible utility in constructing the fields. Results obtained are of interest for communication with, or detection of, objects located in the refraction shadow region of an antenna illuminating an inhomogeneous lossless environment, and for similar problems involving lossy media; for gap coupling to totally-reflected optical beams; and for calculation of diffraction losses in open optical resonators.

Alternative to the complex ray approach is a new theory, which provides a means of tracking inhomogeneous wave fields in real coordinate space. Examples of inhomogeneous wave fields are evanescent fields, leaky waves and Gaussian beams. The advantage of this theory is that it provides direct information on phase front distortion and amplitude changes as the wave or beam penetrates a medium and is reflected,

refracted or scattered; in particular, the theory accommodates scattering events that lead to substantial field distortions as, for example, scattering of a Gaussian beam by an obstacle whose size is comparable to the beam width. Results have been obtained for various evanescent wave and beam scattering problems, and further studies are in progress. A very recent development involves application of evanescent wave tracking to modal propagation in graded-index optical films and fibers. This entirely new analysis promises to render tractable a broader class of index variations than can be accommodated by presently used modal or ray-optical techniques.

It has long been known that there are various biological hazards attendant on the use of high microwave power. The human is particularly sensitive in this respect. There exists, therefore, strong motivation for the establishment of tolerance limits and, ultimately, safety standards. A study of the effects of microwave radiation on the eye, employing rabbits as subjects, is continuing.

B. ACOUSTICS

Program Director: H. L. Bertoni

A major program that has been underway for several years deals with guided acoustic waves propagating on the surfaces of, or at interfaces between, elastic solids. Because of the extremely low velocity of elastic waves, as compared to electromagnetic waves, and their correspondingly small wavelength, these waves have found application in a variety of miniaturized signal processing devices, whose electromagnetic counterparts would be cumbersome, or not at all feasible. Such devices include delay lines, pulse compression filters, band pass filters, high frequency resonators and, in conjunction with semiconductors, convolvers, correlators and storage elements. These devices involve acoustic wave interactions and are therefore acoustic in nature, but they are fitted with input and output electro-acoustic transducers and are employed in electronic systems.

Motivated by these applications, the program has sought to explore device applications as well as the basic wave scattering and coupling phenomena that occur in such devices. Many of these scattering and coupling phenomena also arise in methods which are used for the nondestructive testing of solid materials for flaws, cracks, etc. Our program is currently exploring the ways in which known methods of nondestructive testing or evaluation can be made more quantitatively accurate, and also devising new methods.

In order to study acoustic wave scattering and coupling phenomena in a systematic manner, we introduced and developed a microwave network and transmission line formalism for acoustic waves in isotropic solids. Using this formulation, transmission-line representations have been derived for bulk acoustic waves, and equivalent networks were obtained for several types of planar interface. These networks in turn were used to derive the characteristics of waves guided by a variety of planar structures, such as free or welded plates and plated surfaces.

The transmission line and network approach also forms the basis for a systematic investigation of the properties of several basic waveguides for acoustic surface waves. These waveguides are linear (as opposed to planar) configurations which serve to laterally confine the acoustic fields; the structures under detailed examination in the program include the strip and slot waveguides, which are flat overlay guides, and the rectangular ridge topographic and tall overlay waveguides. The studies are primarily theoretical, but some measurements are also being taken. The analytical results for the strip and slot waveguides are the most accurate ones available, and the analyses

for the rectangular ridge structures form the only analytical theories to date for those guides, although very accurate numerical results have been published for the topographic structure. These analytical expressions, furthermore, agree well with both our own measurements and those of others. We are also studying the properties of the ridge structure of semi-infinite height, which can guide waves along its top edge, in order to better understand the behavior of ridge waveguides of finite height.

The transmission line and network approach is not only useful in the analysis of waveguides, but it has also permitted quantitative design in a novel method for the efficient excitation of these linear waveguides. This method is a two-dimensional version of the prism coupler used in integrated optics for coupling a laser beam to a surface wave on a thin film. In its application to the excitation of the acoustic strip waveguide, an efficiency of 65% over a 25% band was achieved without any cut and try.

Novel components for acoustic surface waves are also being devised, and they are being analyzed with the help of the networks mentioned above. Among these components are simple filters, bulk wave suppressors, power splitters, nonreflecting delay-line taps, and resonant cavities, all based on either total reflection or partial reflection produced by a narrow strip of plated material. These components have the added advantage that they can be used on any substrate, piezoelectric or not. The effect of a beam of finite width on the performance of these components is presently under theoretical study. A transmission line representation for bulk waves in piezoelectric crystals has also been developed, and it was used to treat a novel class of bulk wave filters.

An important class of surface acoustic wave devices for signal processing employs nonlinear space charge effects in semiconductors adjacent to piezoelectric materials. Studies being performed here relating to this class of devices include convolvers, an acoustic phase-locked loop, correlators, time compressors and storage devices. The use of semiconductors in proximity to a piezoelectric substrate was first proposed at the Polytechnic; subsequent work here has added to both the theoretical and experimental development of such devices. Besides the original work on convolvers, the acoustic phase-locked loop was invented here, and a fully integrated version is now under development. Real time correlation and time compression were also obtained here, and these devices are currently being refined, together with a storage correlator.

The first complete theoretical analysis has been carried out of the phenomena associated with the reflection of a bounded acoustic beam from the surface of a solid at the angle for phase matching with the surface wave. The primary effect is a large lateral shift of the reflected beam from the position predicted by geometrical optics. Experimental observations carried out elsewhere for the case of a beam incident from a fluid are in complete agreement with the theory. Subsequent studies for a beam incident from a second solid have led to a complete theoretical understanding of the performance of wedge transducers for exciting surface waves; this theory has shown, for the first time, how high efficiencies can be achieved for such a device. These phenomena also offer a new method for the nondestructive evaluation of the surface properties of a solid, and their use is currently under investigation.

In addition, we are studying the propagation of waves resulting from the coupling of magnetic moments in a ferrite material to the elastic stress. The dispersion and polarization characteristics of such magnetoelastic waves have been found for both bulk and surface waves.

C. OPTICS

Program Director: L. Bergstein

Great progress has been made recently in broadening the research program in optics. The theoretical studies have been extended and the experimental effort has been expanded to include the fabrication and measurement of integrated optics structures and related thin-film devices. The present program consists of the following topics: the development of a new narrow-band matched optical filter (theoretical and experimental), research in integrated optics and related thin-film devices (theoretical and experimental), and the investigation of unstable optical resonators (theoretical only). The topic of integrated optics is the one which is expanding most rapidly, and it currently consists of several different projects.

Development is progressing on a class of narrow-band optical passband and/or rejection filters based on selective reflection from atomic vapors. These filters, which can operate in the 580 to 8550Å range, are rugged, of small size, and have a large acceptance angle and high spectral resolution. Because of these factors, the vapor filters are expected to be superior to presently available devices in pollution monitoring and other applications which require simultaneous multichannel spectrochemical analysis for trace elements. They are also expected to be useful in the detection of isotope concentrations, as a narrow-bandwidth ultraviolet mirror for high intensity radiation, and in astronomical observations. Tests performed on a mercury vapor filter verified the theoretical predictions. With a center wavelength of 2537Å, the filter has shown a peak reflectance of 40%, an acceptance angle of 7 degrees, and a bandwidth of only 0.14Å. The filter was used in an emission flame photometer to perform a spectrochemical analysis for mercury in a water solution. The device was also used to take a monochromatic photograph of the spatial distribution of excited mercury in the flame from which an estimate of the flame temperature profile could be made.

This device principle is being extended to elements which are more difficult to use since they require a higher operating temperature. An example of this type of element is sodium. If a successful filter can be built using sodium, the same structure can be used to contain a variety of elements. At present the fabrication of a cell utilizing a nickel housing and sapphire windows is nearing completion.

The purpose of the program on integrated optics is to develop film-layered structures and other micro-optic devices for eventual application in optical communications. It began with an investigation of the coupling and progression of optical beams along and through thin films and other optical waveguides. Planar and cylindrical polarization-independent waveguides have been investigated and were shown to be practical. Such waveguides can support non-polarized beams and can transmit information at twice the rate of ordinary waveguides. For planar waveguides it was found that phase and group velocity matching of a TE and a TM signal can be achieved by the addition of only a single thin-film layer. Such a scheme can also be used in nonlinear interaction applications to match the group velocity of two optical signals of different wavelengths and/or field distributions. Results analogous to those found for planar waveguides have also been found for cladded fibers. An approximate (but general) transmission line equivalent which we have developed for this class of structures reduces considerably the analytical complexity of the problem and allows the determination of the modal dispersion characteristics in a fairly straightforward manner. Moreover, we have found a realizable radially-variable refractive index distribution n(r) for which all rotationally-symmetrical TE and TM modes have the same group velocity and spatial power distribution. The modes and eigenvalues of such an optical fiber and its dispersion properties were found.

The beam-shift phenomenon which occurs when a beam is incident on a planar structure at the angle corresponding to a leaky wave supportable by the structure was mentioned in the section on Electromagnetics. This phenomenon was investigated theoretically in quite general terms, and the conditions for optimum beam-to-surface-wave coupling due to optical beams of finite width were derived for both uniform thin-film couplers and grating couplers.

The techniques developed by us previously for treating periodic structures have been combined with the beam-shift considerations and applied to the detailed analysis and design of planar grating couplers. These integrated-optics components serve the important function of coupling an optical beam into (or out of) a surface wave guided by thin films. In particular, an exact solution of the boundary-value problem posed by dielectric gratings has been formulated and a computer program has been set up to generate highly accurate numerical results. In this manner, the coupling efficiency, operating conditions and other data have been determined for a large variety of grating couplers. In addition, a simplified perturbation procedure requiring much shorter computation times has also been developed and found to yield excellent approximations of the exact results. These techniques are being used to determine dimensions for a class of blazed gratings that are capable of a substantially higher coupling efficiency than that of ordinary gratings. All of these techniques are instrumental in providing a range of simple criteria for the design of dielectric gratings having desirable diffraction and waveguiding properties.

A laboratory facility has been established and experiments are now in progress to verify some of the analytical results and to assess the performance of certain micro-optic devices. The facility includes a capability to produce solution-deposited films with a view towards making active devices, and a scanning electron microscope which has been modified to write patterns for optical waveguides, gratings and circuits. A computer-controlled digital interface was developed to generate and control the desired scanning pattern.

A model has been developed to describe the exposure and development characteristics of photoresist. The accuracy of this model has been verified and the model has been applied to the fabrication of grating couplers. The couplers so constructed were found to have properties that agree well with the predictions of the theory mentioned above.

Several additional theoretical investigations have recently been initiated relating to the area of integrated optics. One of these is an extension of the beam-shift considerations, mentioned in the Electromagnetics section, involving lateral waves (rather than leaky waves, which are appropriate to the grating couplers mentioned above). These lateral ray and beam shifts on boundaries with incidence-angle-dependent reflection coefficients are currently receiving special attention because it has been recognized that the correct ray-optical model for modal propagation in films and layers requires laterally shifted paths. A systematic study of lateral shifts, both longitudinal (Goos-Hänchen type) and transverse (Imbert type), for isotropic and anisotropic, homogeneous and graded-index, two-dimensional and three-dimensional, guiding regions is presently in progress.

Another new theoretical area involves the determination of the propagation characteristics of linear waveguiding structures for integrated optics. Under investigation are two waveguides which are composed of metal-clad regions on thin films located on a substrate. One of these guides consists of a slot between two metal-clad regions; a basic constituent mode was neglected in analyses performed elsewhere, and we find that important performance differences occur in a certain frequency range.

The other waveguide is a novel one consisting of a metal strip on a thin film on a substrate. This guide employs the "plasmon" mode and is lossier than other known linear waveguides, but it seems to be dispersionless over a large frequency range and may therefore be of distinct interest. The properties of these two guides are being examined in greater detail.

An additional theoretical topic is an application of the technique for the tracking of inhomogeneous fields, mentioned in the Electromagnetics section. As indicated there, application is being made to modal propagation in graded-index optical films and fibers; this new approach permits the treatment of a wider class of refractive index variations than can be handled by other techniques.

Several novel thin-film devices have been devised and investigated both theoretically and experimentally. Thin-film polarizers, beam splitters, and (fixed and tunable) narrow-bandwidth spectral filters for the visible and 10-micron region have been investigated and were shown to be feasible in both the usual optical beam-shaping applications and in integrated optics applications, where these signal shaping components can become part of the coupling scheme. Experimental tests have confirmed the theoretical results. A five-layer FTR narrow-bandwidth spectral filter and a single-layer FTR beamsplitter with transmission properties which are independent of the polarization of the incident beam were developed and tested and their properties were found to agree with the theoretical predictions.

Another new project involves the analysis of certain geometrical discontinuity problems. As a first example, we have solved an important discontinuity problem for a class of planar optical waveguides; the discontinuity consists of a change in height of a dielectric layer. In our modal expansion approach, which is fairly general, we place the open waveguide structure between two conducting planes and then allow the separation between the planes to reach infinity. This greatly reduces the complexity of the problem and permits the solution of a wider class of problems than is possible with other methods found in the literature. The solution we have found for single-mode waveguides agrees well with those found in the literature by less general and/or more cumbersome methods.

The investigation of the properties of unstable resonators is continuing. Two independent approaches are employed; one involves a longitudinal approach whereas the other uses a transverse, or waveguide, viewpoint, discussed briefly in the section on Electromagnetics. Using the longitudinal approach, a meaningful, approximate solution has been found for the complete set of modes and associated eigenvalues of optical resonators filled with an active medium having transverse variations in both the gain and refractive index. This has led to a simple physical picture which permits substantial new insight into the mechanism of operation of such resonators and will help in the determination of the optimum resonator parameters which are to be used in high-power laser devices.

For unstable resonators whose transverse boundaries are determined by the mirror dimensions, the waveguide approach to unstable resonators formed by hyperbolic strip mirrors has been found successful in explaining in a remarkably simple manner the complicated behavior of the complex eigenvalues of the resonant modes. These results are now being extended to circular mirror configurations and attention is being given to the effects of inhomogeneity in the resonator medium.

D. QUANTUM ELECTRONICS

Program Director: W. T. Walter

The research program in the area of Quantum Electronics comprises that portion of electronics in which quantum mechanical effects become important because of the discrete nature of charge carriers, matter or radiation. Optics and Wave-Matter Interactions are related research areas whose boundaries with Quantum Electronics are not distinct. Some of the research previously reported in the Quantum Electronics area has been shifted in this report to the Optics or Wave-Matter Interaction research areas.

A research program has been developed which has attacked outstanding problems of quantum electronics on a broad front. Significant results include: the development of a new type of optical and infrared spectroscopy which has demonstrated more than two orders of magnitude improvement in resolution compared to previous techniques, a corresponding improvement in the long-term frequency stabilization of lasers, significant progress in the understanding of the irreducible quantum fluctuations which are present in the output of lasers and other quantum electronic devices, advances in the understanding of the propagation of ultra-short pulses in resonant media, progress in the development of a new wide-angle low-noise laser receiver, an improved understanding of the properties of the acousto-optical interaction and the use of this interaction to provide acoustical gain and also to mode lock a ruby laser, a significant improvement in the brightness and in the average power generation density of the very-high-gain pulsed copper vapor laser, construction of a computer model to analyze these pulsed metal vapor lasers, and the development of a procedure for calculating improved values of atomic transition probabilities.

One group of projects has centered on the study of the nonlinear response of saturable absorbers to intense laser beams. These projects have yielded a new type of ultra-high-resolution spectroscopy, significant improvement in the long-term frequency stabilization of lasers, and have also contributed to the progress on an automatically-adapting filter which offers the promise of a significant improvement in the performance of an optical radar receiver. Laser Saturated Resonance Spectroscopy (LSRS) has been applied to the study of the absorption spectrum of the SF_6 molecule at 10.6μm and the I_2 and Br_2 molecules at 0.6328μm. A resolving power of 10^8 was demonstrated, representing a two order of magnitude improvement over conventional spectroscopic techniques. The study of the LSRS of xenon has now been completed. All of the observed components in the LSRS spectrum of a naturally occurring mixture of xenon isotopes (masses 128, 129, 130, 132, 134 and 136) at 3.5μm have been resolved and identified. The narrow Lamb dips have also been used to stabilize the frequencies of both the CO_2 (10.6 μm) and He-Ne (0.6328 μm) lasers. A laser which is frequency stabilized to a Lamb dip is a strong candidate for selection as an optical (or infrared) wavelength standard, and will also be useful for long-range interferometry, for deep-space communication links, for coherent optical radar, and for seismic studies.

The high intensities achievable in laser beams can perturb atomic energy levels and radically alter the position and shape of spectral lines. The theory of these phenomena is being extended and refined. Recent studies have centered on the inclusion of finite lifetime and recoil effects. New results were obtained on the effect on the molecular center-of-mass motion of a resonant driving field. The analysis exploited a novel calculation technique based on the application of Floquet's theorem to periodically-driven quantum-mechanical systems. It is hoped that these interesting effects at optical frequencies will eventually be observed and compared with theory.

Another theoretical investigation in progress aims at understanding the irreducible quantum fluctuations always present in the output of lasers and other quantum electronic devices, from the point of view of first principles. A study has been completed of the effect of an internal homogeneously-broadened saturable absorber on the intrinsic fluctuations of a laser oscillator. A related effort is concerned with the understanding of spontaneous emission processes in resonant dispersive configurations other than free space. This problem has been solved in general terms by relating it to the problem of calculating thermal field distributions. Recently, the spontaneous emission noise and the response characteristics of driven quantum electronic systems have been evaluated for classes of spatially degenerate molecules. Effects associated with the nonstationary character of the correlation functions have been particularly emphasized. These effects have been overlooked in recently published work on the response of driven systems.

The study of the propagation of ultra-short pulses in resonant media has led to the discovery of new relationships between their spatial and temporal structures. These include: strong self focusing, phase modulation which accompanies transverse confinement of the pulse, additional unsuspected structure of plane wave pulses, some effects implied by the inclusion of a finite transverse relaxation time, the transverse stability of confined fields, new types of distortionless fields which simultaneously possess phase modulation, polarization modulation and three-dimensional structure, and some effects produced by a nonlinear index of refraction of the host material.

The spectroscopic investigations of saturable absorbers and the study of propagation of ultra-short pulses in resonant media relate directly to the feasibility study of the low-noise wide-angle BALAD (Bleachable Absorber Laser Amplifier-Detector) receiver. In this receiver, an absorption cell (filled for example with SF_6) is placed between the laser amplifier and the detector to serve as a threshold filter, both in the frequency domain and in two spatial dimensions, and it screens out all background light which is not coherent with the signal beam. Calculations predict superior performance in optical radar and communications applications. The LSRS studies have clarified the effect on laser amplifier gain, bandwidth and band center that would result from the use of separated single xenon isotope samples in an optical radar system with a BALAD receiver.

Visible and near-visible lasers with an efficiency and power comparable to the far infrared CO_2 laser are still lacking. The vapors of certain metals and transition elements have attractive energy level structures. A number of high-temperature laser systems have been constructed to explore the most-promising metal vapor laser media. The pulsed copper vapor laser has high gain and high peak power at 5106Å. Moreover, unlike other self-terminating lasers, single mode output and therefore high peak brightness can be obtained. Recent experiments using higher pulse repetition rates and a larger volume have produced an average power of 11 watts. With argon as an additive gas, average powers above one watt have been achieved at 5106Å from an active region only 10 cm long and 1.25 cm in diameter, corresponding to an average power production of 0.1 watts per cubic centimeter. A heat-pipe copper vapor laser has been demonstrated, and preliminary experiments with a high-loss resonator have yielded a peak brightness of 10^{12} watts per square centimeter per steradian, which is the highest brightness measured for a gas laser. Heat-pipes and the use of dissipated discharge energy for self-heating have potential for practical sealed-off copper vapor lasers with average output powers up to 100 watts in the visible. Still higher output powers and overall electrical efficiencies of up to 10% are anticipated from the current investigation of impedance tailoring of copper vapor systems. Other metal vapor systems are being examined for efficient laser action in the ultraviolet and blue-green spectral regions. A computer model has been constructed and is being used in conjunction with the experimental analysis of several metal systems. Theoretical analysis of these systems for higher efficiency lasers has led to the development of a procedure for calculating improved values of atomic transition probabilities from spectral intensities.

E. SOLID STATE AND MATERIALS

Program Director: H. J. Juretschke

The solid state research ranges over a wide area of experimental activities encompassing the creation of new or better materials, the investigation of many specific properties of solids, and the application of these materials and effects to practical situations. In addition, there are active programs in various aspects of solid state theory.

In the area of materials development, an important sector is concerned with new magnetic compounds involving rare earth and transition metal ions, and with the production of phosphors for IR-to-visible conversion. Programs involving thin films include the development of high-mobility semiconducting layers, and the preparation of highest quality single crystal films of the noble and ferromagnetic metals and the semimetals. In addition, there are systematic efforts to produce good insulating barriers on semiconductor and metal single crystal surfaces, for application to tunneling and other devices.

Among the properties of interest, those related to electromagnetic interactions play a central role. Electron transport in semiconductors is studied for hot-electron nonlinear effects, and for surface interactions. Size and surface effects in carrier transport in magnetic fields, especially as related to the band structure and to surface relaxation mechanisms, are investigated in thin films. One group of studies centers on the spin dependence of electrical conduction in magnetic materials, while another is concerned with the characteristics of thin tunneling junctions, either vacuum or insulating, between normal or superconducting materials. New research includes the interaction of x-rays with electrons in semiconductors through the internal photoelectric effect, the determination of charge concentrations at metal-insulator interfaces by surface electron scattering, the details of metallic surface self diffusion as it affects the electrical surface properties, and the concentration profiles in alloys of thin films as affected by the large surface-to-volume ratio.

Studies in magnetism include a study of the magnetic coupling between transition metal ions in complex fluorides, using neutron diffraction, magnetic susceptibility and Mössbauer spectroscopy. Magnetic resonance measurements (NMR, ESR) are being used to study clustering of rare-earth ions doped in CdF_2. This material has been developed as a base for infrared to visible light conversion, which has potential application in a variety of display devices. Another study is looking at the origins of the frequency dependence of magnetic anisotropy in ferrites and at the anisotropy of their magnetogyric ratio. A program in magnetic resonance of ESR and NMR is elucidating details of electronic structure in many of the materials used in other investigations.

Optical investigations include a broad investigation of the Faraday effect in ferromagnetic metals that aims at clarifying details of their band structure, especially as a function of temperature. At very short wavelengths, in the x-ray region, studies are pursued on quantitative aspects of the resonant Borrmann transmission in perfect single crystals under conditions where three or more distinct beams are in strong interaction. Here an analysis of loss mechanisms has been critical, and we have initiated work on better determination of x-ray dispersion coefficients, and on the energy distribution of the propagating waves relative to the atomic sites.

X-ray investigations are concerned with crystal structures, perfection, thermal effects and charge distributions in a wide range of materials. Determination of lattice constants through multiple diffraction measurements are yielding data of considerably enhanced accuracy and precision. Multiple diffraction effects are also being applied in a careful interpretation of x-ray data, including linear absorption coefficients, line shapes and line shifts.

Theoretical studies include a fundamental program in electron-photon interactions in magnetic fields, and the theory of slow-electron scattering in solid surfaces. In addition, there is work on elastic and inelastic interactions between electromagnetic radiation and atoms and solids involving excitation of inner electrons, especially as they apply in the x-ray region and on the phase-sensitive coupling of optical fields in solids to suppress noise. Finally, the nonlinear coupling between x-rays and other waves is studied in the domain of dynamic x-ray modes.

Another, and new, aspect of solid state studies involves the theoretical and experimental investigation of distributed parameter networks on integrated electronic circuits. There has always been a search for ways to utilize an existing technology to its fullest advantage. By using a distributed parameter viewpoint on silicon planar technology, it is possible to synthesize functions on chips which are much smaller and less expensive than chips containing lumped elements only. The result of initial tests have shown the feasibility of manufacturing totally-integrated voltage-tunable RF amplifiers, IF strips, FM demodulators, oscillators, and other frequency-selective electronic networks. The elimination of large lumped inductors itself results in an appreciable saving in size, weight, material, and ultimately, cost. The design is simple and well suited to mass production techniques in silicon technology.

F. WAVE-MATTER INTERACTIONS

Program Director: N. Marcuvitz

Several different programs comprise the area of wave-matter interactions. The most comprehensive and challenging of these programs is a relatively new one which involves the interaction between very high power electromagnetic waves and materials, in which the interactions are highly nonlinear and turbulent. Other programs have been underway which treat interactions appropriate to more moderate energy levels; such interactions involve quasi-linear processes which can be treated either by equations possessing weak nonlinearity or by linearized equations, the latter leading to a large variety of parametric processes which have been studied by us in detail over a period of time. The above-mentioned studies are primarily theoretical but they have experimental phases as well. Other programs include various experimental and theoretical investigations of plasma wave properties and plasma turbulence effects, and studies in space radiophysics, which include terrestrial and extraterrestrial phenomena.

A comprehensive program has been initiated to study phenomena associated with the interaction of high power density electromagnetic waves and materials. Its motivation is to extend known techniques on linear wave propagation in spatially inhomogeneous and time varying deterministic media to practical structures wherein the material and waves are nonlinear, inhomogeneous, time varying, and turbulent. Of particular interest are self-consistent studies of reflection, transmission, and mode conversion properties of high power waves as a function of frequency, pulse width (spatial and temporal), etc., for the range of power densities wherein material properties become markedly nonlinear and experience phase changes. Preliminary model computer studies are being carried out with the aid of an interactive graphic interpretive (IGI) language developed for a PDP 11 computer facility; related experimental and theoretical investigations are also under study.

Another program, which is also relatively recent, involves weakly-nonlinear interactive systems; the approach adopted employs a scattering formulation applied to nonlinear transmission lines. A recent result, involving the general problem of reducing a nonlinear equation to a linear one, has been to construct, in a simple and straightforward manner, a general such transformation for a class of generic equations, of which

the well-known Burger's equation and the K-deV equation are special cases. Another important result has been to show on a rigorous basis that the energy in a lossless linear transmission line is conserved, in contradiction to the published literature which characterizes the presumed violation of the conservation law as a paradox.

The program on parametric processes, which involve the linearization of weakly-nonlinear interactions, has been highly successful over an extended period and now includes application to nonlinear optics and to processes involving plasmas. Recent theoretical studies have stressed the interaction of light with plasmas, including the study of stimulated Brillouin scattering (SBS) and parametric-decay instabilities. Each of these nonlinear-optical interactions involves the ion-acoustic wave in the plasma. The dependences of the instabilities on the dispersion, absorption, anisotropy and inhomogeneity of the medium are of particular concern. The results of this investigation have practical applications in laser-produced plasmas for fusion and ionospheric probing by radio waves. Other applications of parametric interactions are found in the development of new coherent optical sources for communications, and in design of pulse-echo devices for data processing. Theoretical studies have employed a Floquet expansion procedure which takes all the space-time harmonics into account, and is rigorous within the small-signal regime. The rigorous boundary-value problem and the initial value problem for instability evolution for an intense light beam incident on a medium are being solved and applied to a variety of physical cases. Instability thresholds in unbounded regions have been further investigated for interesting cases where two interactions (involving three frequencies) merge into one interaction (involving four frequencies) upon application of a small change in model parameters. Such merger of interactions appears likely in certain parametric interactions taking place in gradient inhomogeneities. The study of a class of problems dealing with intracavity parametric interactions has been initiated, prompted by interesting experimental results obtained in mode locking a ruby laser using stimulated Brillouin scattering in a birefringent crystal. In other experiments, the dynamic build-ups of SRS and SBS phonons in crystals are being studied on a time and space resolved basis in order to illustrate some of the theoretical predictions mentioned above.

The research activities related to plasmas include the examination of plasma waves and the study of plasma turbulence. Both theoretical investigations and laboratory experiments are being conducted in these areas. Both linear properties and nonlinear effects of plasma waves have been and are being examined, and related experiments are being performed to verify the theoretical findings. Among the topics which have been treated are: a) the propagation characteristics of electrostatic plasma waves in a magneto-plasma; b) plasma heating using the resonance damping of cyclotron waves; c) evolution of parametrically-excited ion-acoustic instabilities and plasma heating; d) propagation of kilovolt-subnanosecond base-band pulses along a plasma column; e) mixing of electromagnetic waves in a plasma through modulation of the electron temperature; f) coupling effects between electromagnetic modes and electrostatic modes; g) examination of the wave properties related to microwave interferometry in finite-size plasmas; and h) resonance excitation of plasma waves by a slotted cavity. The effort in this area is being continued and coordinated with the study of plasma turbulence.

The study of plasma turbulence has been gaining more emphasis recently. A comprehensive theoretical restructuring of plasma turbulence has been carried out, and new experimental facilities were designed and built to perform sophisticated experiments on plasma turbulence. One of the new experimental facilities is a long, hollow-cathode arc system, which can produce both quiescent and weakly-turbulent highly-ionized plasmas. Not only were the static characteristics of weak plasma turbulence studied, but the dynamic effects of the transition from the quiescent state to the weakly-turbulent states were also observed. Advanced data processing techniques were developed to gain reliable data on plasma fluctuations. One of the distinguished achievements in this area

is the clear quantitative evidence of the plasma turbulence generated by the electrostatic ion-cyclotron waves and of the turbulent diffusion process. Dynamic or feedback stabilization techniques are being applied to isolate the various terms contributing to the growth rate of the instability and the transition into the weak turbulence regime.

Associated with the investigation on plasma physics is the research in space radiophysics of natural terrestrial and extraterrestrial phenomena. Emphasis is on the physics of the upper atmospheres and ionospheres of the planets and their interaction with the sun's emanations. The program involves theoretical and experimental studies of disturbances in the ionosphere as produced by man-made events and by natural phenomena, such as thunderstorms. Our facilities permit the sounding of the ionosphere for disturbance detection. Radio emissions at VLF as detected by satellites have also been studied. The problem of the coupling between waves in the neutral atmosphere and waves in the ionosphere is presently being studied. Resonances have been discovered and it is planned to seek confirmation both from spacecraft data and from ground-based measurements. Dynamic interaction of the atmosphere and ionosphere of planets due to heating by solar radiation has also been studied from the standpoint of the outflow of gases and ionization as related to the formation of the solar system. These methods have been applied to the Saturnian Satellite Titan and to Jovian satellites, which are unique in the Solar System.

G. ELECTRIC POWER ENGINEERING

Program Director: E. Levi

The recent energy problems and the increased awareness of the deterioration of the environment are likely to bring about a significant increase in the share of energy utilized in electric form. In response to these mentioned needs, the Polytechnic has initiated a new research program, and has enriched its academic curriculum by instituting electric power engineering options at both the undergraduate and graduate levels.

Under investigation are many critical areas in the generation, transmission, and distribution of electrical power. Of particular interest are: (1) the magnetic separation of radioactive isotopes in nuclear fuels and waste products; (2) the application of novel technologies to the development of individual components, such as generators, high voltage transmission lines, gas insulated substations, circuit breakers, fault-current limiters, rectifiers and inverters; (3) the stability, reliability, and economy of large integrated systems; and (4) their effect on the environment and the quality of life.

At the utilization end, transportation has been singled out as the primary area of concern, since it consumes about one quarter of the overall U. S. energy budget and one half of the oil, and since it creates serious pollution problems. In the U. S. only one percent of the railroad is electrified, as compared with more than 90 percent in the other industrialized countries of the world. Besides seeking improvement in passenger and freight train services, new modes for individual, as well as mass transportation are being developed. These include electrified urban thoroughfares, as well as highways. A significant effort is devoted to linear propulsion by means of iron-cored synchronous motors. These motors present challenging problems in electromagnetics because of their complex geometries; a recent study involves a new method for the determination of magnetic fields in air gaps. Another problem which arises in such linear motors is the need to achieve a more accurate assessment of eddy current losses and field penetration in thickly-laminated iron structures. In addition, as a result of pioneering work conducted at the Polytechnic, novel topologies for flux inter-linkages have been conceived.

II. SYSTEMS

A. COMMUNICATIONS

Program Director: L. Kurz

The research program in communications and information processing currently covers topics in the optimization and evaluation of digital data transmission systems, robust detection and estimation techniques, and FM receivers.

The transmission of digital data in the presence of intersymbol interference and noise is a classical problem in communications. The problem has recently taken on an even more important significance as data rates become increasingly higher and, therefore, the effects of intersymbol interference and noise more severe. The main research effort in this area was concentrated on the optimum and suboptimum signalling and detection in the presence of intersymbol interference and gaussian noise, detection in the presence of noise represented by a mixture distribution model, and techniques useful in the monitoring of channel conditions. In particular, a new design procedure for recursive minimum mean-square error equalizers was developed, two types of detectors using digital filters were investigated, suboptimum detectors for signals corrupted by gaussian and impulsive noise were analyzed, and two classes of direct and indirect robust estimators of channel conditions were introduced and compared to existing procedures. Work is continuing on all the above-described areas of digital data system optimization.

Considerable interest generated in detection, feature extraction and estimation problems, where little is known about the data to be processed, motivated the development in recent years of several classes of rank and non-rank robust (insensitive to underlying distributions) detection and estimation procedures. Though the rank detectors tend to be more powerful, its non-rank competitors are more robust and simpler to implement. Thus, the main research effort was concentrated on developing further the latter class of detectors stressing applications in the detection of non-constant signals in mixture distributions and fading environments, sequential and nonsequential detection of underwater sounding and two-dimensional data. A parallel effort included the development of robustized estimation techniques which are applicable in adaptive modes of operation in data processing. In particular, a new family of variable-threshold non-rank procedures was introduced and compared to competitors based on the rank procedures; slippage and analysis of variance techniques were modified to operate on an adaptive mode, a new class of quadratic three-sample partition detectors, which is useful on change-of-scale and stochastic ordering detection problems, was generated and shown to be an effective and easy-to-implement class of competitors to the rank detectors; various techniques were developed for finding optimum scores and thresholds for partition detectors; two basic theorems pertaining to robustized recursive estimation were proven. Work in all the areas of robust detection and estimation described above will be continued.

Another study is concerned with investigating a novel FM demodulator capable of combatting the effects of interfering signals. This demodulator, consisting of two phase-locked loops interconnected in a feedback arrangement, has been constructed and is being studied. Further theoretical and experimental analysis will continue.

B. COMPUTERS AND COMPUTER-COMMUNICATION NETWORKS

Program Director: E. J. Smith

The research program in Computers and Computer-Communication Networks covers a number of topics directed toward the development of improved computer architecture, improved languages or information structures, techniques for implementing algorithms, data networks, and message switching systems.

In computer communications, we are concerned primarily with techniques for the better understanding of and improved design of data and computer networks. The stress is on message store-and-forward networks with minicomputers used to combine or concentrate incoming messages, to buffer or store them if need be, to route messages either dynamically, or following prescribed routing algorithms. There are interesting questions here of the appropriate modeling of interconnected networks of computers to carry out the necessary analytical work and of the appropriate choice of message statistics to be used in the modeling. The combining or concentration function is being studied, with comparisons made between different combining techniques. Dynamic buffer schemes are being studied, with optimum buffer size and comparison of various schemes a specified goal. From the overall network viewpoint a comparison of various adaptive routing algorithms is underway, as well as methods for maintaining flow control throughout the network. Also being studied are algorithms for network design taking into consideration topology, capacities, routing, reliability, and other factors.

Three specific projects have been completed which deal with adaptive routing and resource allocation, multiple routing networks, and resource-sharing in computer-communication nodes. The present effort focuses in particular upon teleprocessing network design algorithms, channel assignment in mobile telecommunication systems, and adaptive message-switching networks.

Investigations of the dynamic behavior of computer-driven communication systems are concerned with the development of models which might be useful as analytical tools in the evaluation of such systems as well as in the ultimate design of improved systems. In order to provide a focus for the effort, a particular message-switching system is studied in detail and the approach is directed toward the development of an interconnectivity model in which the system is viewed as a collection of program modules, each of which communicates with other modules through an interconnection medium via a particular machine. This often obscures the original intent of the designer of the system and provides no insights into the necessary data structures and operations for performing message-switching functions. Motivated by these considerations, current work is directed toward the development of a higher-level algorithmic description language for use in the design and specification of switched communication systems. Desirable characteristics of such a language are defined and, as an initial step, the ideas are applied to the description of a communications scanner channel. The modules are characterized by various parameters such as core size, execution time, data base, communication linkages to other modules, etc.; and the application of the model is viewed with respect to the evaluation of system changes and debugging difficulty, throughput, and local optimization of individual models. A first version of a communications-oriented design language has been completed and a current effort is concerned with the investigation of appropriate data structural forms for efficient implementation of the language. Also recently completed was an approximate analytic model of a message-switching system which provides the capability of predicting message-processing delays from a knowledge of the message switch architecture and the traffic statistics, as well as permitting one to estimate the effect on performance of transferring a task from one processor to the other in a dual-processor system. The current effort attempts to extend the model to a multiprocessor system consisting of microprocessor modules.

In the area of switching theory, a dynamic fault-test generation scheme has been developed for combinatorial logic circuits, and an extension of the same technique is being applied to the case of an asynchronous sequential machine in which tests are generated from the acyclic circuit after a near-minimum number of feedback cuts have been made in the original circuit. Dynamic tests are sought rather than static tests since the former do not require checking and consequently require less complex computation. Techniques are also being investigated for the determination of partitions having the substitution property, or other properties of special interest, through use of the predicate calculus; efficient algorations for computation are sought.

In the area of machine architecture, several studies are in progress, including the exploration of a simple, low-cost, bus-oriented computer having a common memory and instruction format for microprograms and programs as an array processor, and the investigation of a variable-precision arithmetic technique in which all numbers are stored internally within the computer as quotients of integers. The resulting effect upon computational accuracy and machine time will be studied; other related schemes will be sought out and explored.

C. SAFETY, RELIABILITY AND SOFTWARE ENGINEERING

Program Director: M. L. Shooman

The areas of Safety, Reliability and Software Engineering encompass a broad spectrum of analytical modeling, experimental systems research, and systems engineering. The underlying thread of cohesion in these areas is the application of modern techniques of probabilistic modeling, statistical measurement and experimentation, and the development of design and optimization techniques within the engineering process. The emphasis is at times on the basic development of the tools and methodology for such studies and often on the application of these new as well as existing techniques to the advancement of the state of knowledge in the given area.

This section describes and reports the work of several diverse groups within the Institute, some who are developing technology in areas outside of electronics, but with direct application to electronic devices and systems, and others who are working within the mainstream of electronics.

To the probabilistic modeler, in the abstract, the concepts of safety, reliability and availability have a unifying thread. The probability of no equipment, human, or procedural failure which endangers a human being is expressed in the safety index. Similarly, if the failure described in the preceding sentence affects system operation so as to cause a failure we are discussing reliability. Clearly, safety and reliability analysis have the same methodology; however, the definitions, implications, and design goals differ significantly. The availability index is a probability which measures the percentage of systems which are up at any specified time and allows one to model and measure the effectiveness of failure-repair dynamics. Maintainability is measured by repair rate, and repair and failure rate combine through the system configuration to determine availability. Lastly, when we turn our attention to software we find that the concepts of system reliability, availability, and maintainability apply well; however, it is important to emphasize the differences between hardware and software in constructing our models.

Activities in the safety area have recently been stimulated and accelerated by the arrival of new funding from the Department of Transportation for the second year of a proposed three-year intermodal study of transportation system safety. This work is an interdisciplinary effort with several participating departments. Clearly, although much

of the transportation vehicle and its guideway involves mechanical systems, a very high and increasing percentage of modern ships, aircraft, rail, and automobile systems involve electronic control, communications, sensing, guidance, and so on. One of the chief methodological tools being applied to this area is the system fault tree. The most important theoretical aspects are its construction, computerization, and collection of probabilistic input data. In this latter area some new techniques of statistical estimation called consensus estimation are being developed. These differ from, but bear resemblance to, the Delphi techniques used in forecasting. Also of very great importance is the modeling of the human operator. In a transportation system it is impossible to divorce the human operator from the control system. Although a great deal has been done to model the human transfer function, researchers have had less success in modeling human errors which vitally affect both reliability and safety. We have enlisted the aid of an experimental psychologist to work with the systems engineers in this area of research and hope that the cross fertilization of ideas will lead to new approaches.

The reliability (including availability and maintainability) area has been a focus of activity for many years. Many topological methods of system analysis have been developed along with approximate bounds which have served as the basic algorithms for many of the computer analysis programs produced in recent years. Also, much has been done in applying Markov models to a wide variety of problems in availability computation. Recent thrusts involve (1) a queueing theory approach to availability and maintainability which allows more complex repair models; (2) modeling of the inspection interval problem (applied to critical systems), so as to optimize the benefits in safety and minimize the cost; and (3) the study of computer-assisted test of electronic systems with regard to design constraints, optimum test points, and overall cost minimization and reliability maximization.

In the area of software engineering, a substantial portion of this effort is supported by the Rome Air Development Center. Activities in this area are focused upon five aspects of software reliability: a) construction of probabilistic models for software errors, which reflect the content and type of error, removal and generation rate, prediction of mean time between failure and reliability; b) measures of computational complexity based upon the algorithm, automata theoretic or graph theoretic complexity, information content and program size; c) relationship of modular and structured programming styles to reliability; d) modeling techniques for program verification and testing including optimal test strategies for nonexhaustive tests; e) models for the comparison of programming languages, comparative study of formal definition languages, and techniques for compiler writing.

The probabilistic modeling effort dealt with the construction of many-state Markov models for the determination of software availability and macro models to predict number of bugs and program reliability. Preliminary models were developed to predict operational software reliability. In the second phase of this work, error-generation terms were added to increase the degree of realism. A new micro approach relates errors and reliability to a functional path decomposition of the program.

Several approaches are being taken to provide measures of complexity, both for the problem and the ensuing software. The initial efforts treated the complexity of a function via recursive function theory. Present efforts are directed toward relating program complexity to established ideas in the fields of automata, communications and information theory. Recent results provide a measure of program length based on Zipf's Law and operator/operand count.

In the area of structured and modular programming two efforts are being pursued. One effort, presently in progress, applies modular programming techniques to the automatic-interactive construction of programs. In addition, the statistical design of an experiment for the objective evaluation of structured vs. nonstructured programs is in the planning phase.

In the program test area, two efforts are in progress. The objective of the first effort is to develop a method which computes the number of tests necessary to verify a computer program. Verification is categorized into several classes, including "exhaustive" at one extreme and "processing of each instruction at least once" at the other. In the second effort, an analytic method is being developed for the selection of data for automatic program testing.

In the area of languages, a new effort has begun to construct a very high level language for writing program specifications. It is envisioned that this language will lie between a high level programming language (FORTRAN, PL/1, etc.) and English prose, but will be concise and yet avoid ambiguity.

D. SYSTEMS, CONTROL AND NETWORKS

Program Director: D. C. Youla

Research activity in the network area encompasses both the classical and modern lumped-distributed domains. In the latter, several significant breakthroughs have been made which are expected to lead to an exact insertion-loss synthesis technique for the design of optimum filters incorporating cascades of equi-delay TEM lines and lumped, lossless two-ports. In particular, it is expected that transformers and filters exhibiting equi-ripple performance in both pass and stop bands can be designed to specifications. In principle, there appears to be no reason why the method cannot be extended to single-moded waveguide networks employing obstacles to produce the lumped discontinuities. An unexpected and important outgrowth of the above study has been the development of a new diagnostic digital technique for probing one-dimensional nonuniform structures. Hopefully, the modeling of dispersion by means of lumped two-ports will enlarge the scope of the method significantly.

A major topic in the control area involves the application of classical ideas to the design of feedback controllers for linear multivariable systems. All efforts on this topic have been completely successful. A totally new frequency-domain technique has been developed for the design of optimal multivariable controllers. The class of problems that fall within the scope of this technique is much broader than that encompassed by the linear, quadratic gaussian approach (LQG). The latter revolves around the idea of Kalman filtering and for this reason is unable to absorb in a natural and straightforward manner many essential practical constraints. The group is presently engaged in work which should result in an effective computer implementation of the optimal controller. The availability of such an algorithm will undoubtedly suggest related simpler suboptimal strategies and open the door to significant industrial applications.

Work on the control of stochastic systems has been focused on situations in which discrete events, such as equipment failures or message arrivals, occur at random times. Recursive optimization equations have been derived in many cases via dynamic programming. Novel techniques have been devised to solve those equations to get optimal decision rules for diverting messages or vehicles to alternate routes in situations with simple arrival and service time probability distributions. In more complex cases, structural properties of the optimal controllers have been derived and specific control algorithms have been compared via simulation.

In the area of reliability applications, both repairman assignment and optimal inspection problems have been solved. The problems studied here have been of a higher order of complexity than those generally formulated by system analysts. In one case, both feedback control for dynamic tracking accuracy and repairman assignment policies were simultaneously optimized for repairable systems. In another, techniques were developed for optimizing inspection schedules when the corresponding tests degrade the lifetimes of the very components whose status is being checked.

Identification and parameter estimation from random data is of central interest in systems applications. In control applications, one must have better knowledge about the plants as well as effective models of certain plant elements or functions for which there are no intrinsic analytical descriptions available. This is especially true when human elements are in the loop and we try to describe neuromuscular systems such as arm and leg control, pupil response, etc. We also require identification and modeling techniques when the system itself is too complex so that its fundamental input-output properties are unknown. This is especially true of large-scale economic systems where dynamic models are obtained in some optimal fashion from available data in order to describe the evolution of the system as a result of changing policies. Finally, even in the case for which dynamic equations describing a system are known or accepted, parameters are often unknown and must be estimated. Identification and estimation techniques have been applied to all of the problems described above. But, to a great extent, the basic feature of previous applications of identification and parameter estimation is that the models of the systems studied have been time invariant, and their parameters have been estimated from statistically stationary data (after trends have been removed). However, there exist important phenomena for which, even after trends have been removed, the data is intrinsically non-stationary, characterized usually by a period of increasing and then decreasing intensities. Such characteristics are found in economic systems, but are even more prevalent in geophysical observations such as meteorological fluctuations and geological phenomena such as earthquake excitations. Motivated by this last application to characterize the statistical properties of earthquake excitations, recent studies have led to new limit theorems for estimating significant parameters in models of non-stationary time series, and has led to a specific approach for modeling earthquake acceleration data. This approach has been implemented on computers and is being applied to study recent earthquake acceleration data obtained from the Western United States. We are just at the beginning in developing techniques for the identification of non-stationary systems and data. Many problems remain to be solved.

E. DATA PROCESSING

Program Director: A. Papoulis

Current studies in data processing include the following aspects of picture (or tabular) processing and spectral estimation: statistical enhancement and extraction techniques for pictorial or tabular data; reduction of diffraction effects in the imaging of coherent and incoherent objects; image distortion resulting from atmospheric turbulence; and statistical analysis in spectral estimation.

Statistical enhancement and extraction techniques for pictorial or tabular data: This research concentrates on developing procedures to present the data in a useful form or to present the data in some way in which factors in the data which are "almost invisible" become plainly visible. The effort was concentrated on factor analysis and masking techniques. The masking operation was taken to mean the process in which a "window" or "mask" is swept across the data either electronically, mathematically,

mechanically, or some combination of these. The motives for selection of a particular mask are simplicity and effectiveness. The preliminary results indicate that both approaches to the enhancement and extraction problems of tabular data are promising, and the research effort utilizing these and related techniques will be continued.

Reduction of diffraction effects in the imaging of coherent and incoherent objects: The known methods of deconvolution in one and two dimensions for analog and digital data are compared for accuracy and computational economy. A new technique is developed for the complete recovery of objects of finite size. The technique is based on an iteration scheme involving only the FFT. It is shown that, in the absence of noise, the iteration converges to the unknown object. The effects of noise and the aliasing errors are determined and it is shown that they can be controlled by early termination of the iteration.

Image distortion resulting from atmospheric turbulence: In the recent literature, a method has been proposed for a dynamic correction of this distortion. The underlying filtering (Poisson noise) and prediction (control delay) problem leads to a two-dimensional Wiener-Hopf equation whose solution is under investigation under various assumptions about the spectrum of the turbulence. The reverse problem of determining the properties of the medium in terms of image of a moving sattelite has also been considered.

Statistical analysis in spectral estimation: The current emphasis is on the statistical analysis of the method of maximum entropy and in comparison with other methods for various special forms of the unknown spectra. An adaptive perturbation scheme is under consideration for updating the estimated spectra as new information is received. The convergence of the scheme to the smoothed spectrum is under examination.

POLYTECHNIC INSTITUTE OF NEW YORK

G. Bugliarello, President
J. J. Conti, Provost
R. J. Cresci, Associate Provost for Research

B. J. Bulkin, Dean of Arts and Sciences
L. B. Felsen, Dean of Engineering
A. B. Giordano, Dean of Graduate Studies

MICROWAVE RESEARCH INSTITUTE DIRECTORATE

A. A. Oliner, Director
J. Fox
S. W. Rosenthal

FACULTY AND RESEARCH STAFF PARTICIPATING IN MICROWAVE RESEARCH INSTITUTE PROGRAMS

DEPARTMENT OF CHEMISTRY

Faculty

E. Pearce: Department Head

E. Banks
J. M. Casper
P. Hoggard
N. C. Petersen

Postdoctoral Fellows

R. Carvallo
M. Greenblatt
C. Kosky
S. Nakajima

Fellows, Trainees and Research Associates

M. Shone
G. Torre

DEPARTMENT OF ELECTRICAL ENGINEERING AND ELECTROPHYSICS

Faculty

T. Tamir: Department Head

M. Adamowicz
L. B. Bergstein
H. L. Bertoni
J. J. Bongiorno, Jr.
R. R. Boorstyn
F. A. Cassara, Jr.
E. S. Cassedy
B. R. Cheo
K. Chung
D. Davids
H. Farber
L. B. Felsen
A. B. Giordano
S. H. Gross
S. Habib
R. A. Haddad
A. Hessel
F. Kozin
L. Kurz
A. E. Laemmel
J. T. LaTourrette
E. Levi
N. Marcuvitz
E. Mishkin
M. C. Newstein
A. A. Oliner
I. Palócz
A. Papoulis
S-T. Peng
S. W. Rosenthal
H. Ruston
P. Sarachik
H. Schachter
B. Senitzky
L. Shaw
J. Shmoys
M. L. Shooman
L. M. Silber
E. J. Smith
F. Stone
L. Strauss
W-C. Wang
D. C. Youla

DEPARTMENT OF ELECTRICAL ENGINEERING AND ELECTROPHYSICS

Research Scientists

N. Solimene W. T. Walter

Research Associates and Instructors

L. Birenbaum E. J. Lancevich B. Rudner

Research Assistants

A. Green Y. L. Liu

Postdoctoral Fellows

S. Choudhary H. J. Eun E. Kunhardt S. Y. Shin

Fellows, Graduate Assistants and Graduate Students

H. Ahn
S. Austin
D. Baggi
A. Benler
P. Bondyapadhyay
S. Brutzkus
A. Camello
C. Chamzas
K. C. Chang
Z-W. Chang
S. D. Chikte
Y. Devadas
C. Hechtman
Y-L. Hou
C. L. Hsu
E. Hu
H-M. Huang
T. Ishihara
N. Kazanjian
P. Kersten
R. Kinasewitz
H. M. Kudyan
G. Kull
S-P. Kuo
T. Lee
T-Z. Lee
E. Lipshitz
A. Livne
P. Lue
F. Mattar
C. S. Yoon
E. Mayer
S. Mohanty
F. Nakajima
V. Nangia
S. Natarjan
M. Petrella
G. Popkin
D. Ramakrishnan
G. Ramanathan
L. Rosenheck
C. Rubenstein
H. Saidi-Ashtiani
Y. Saito
G. Salayka
C. Santana
A. Sarris
V. S. Shah
N. Shiffeldrim
S. Siddiqi
S. Sinkar
A. Sohraby
J. Sokol
J. K. Stark
F. Templeman
B. Tomasic
N. Wattanapanom
M. Wernicki
K. C. Whang
M. H. Wu
T. Q. Yip

DEPARTMENT OF PHYSICS

Faculty

H. J. Juretschke: Department Head

W. Kiszenick
B. Post
D. B. Scarl
H. W. Schleuning

Research Scientist

S. R. Barone

Research Associate

M. Eschwei

DEPARTMENT OF PHYSICS

Fellows, Trainees and Graduate Assistants

S. L. Chang	H. Grossman	D. J. Lischner	P. Persans
R. DeSerio	J. Gutreich	P. Mazumdar	J. Selby
L. Goldstein	T. Hom	G. Osche	S. M. Turner
	P. Wang		

MICROWAVE RESEARCH INSTITUTE PUBLICATION STAFF

MRI PROCEEDINGS

Editorial: J. Fox, Editor W. Peter, Assistant Editor

MRI TECHNICAL REPORTS

Annual Report Coordination:	J. Fox		J. Maher
Editorial:	J. Maher*		S. Mahrous
Technical Typing:	J. Maher		T. A. Connolly
Graphic Arts:	E. J. Cassano	R. Gnaffo*	J. Morrison

*Principal contributors

PRECEDING PAGE BLANK-NOT FILMED

I. ELECTROPHYSICS

A. ELECTROMAGNETICS

Program Director: A. Hessel

B. ACOUSTICS

Program Director: H. L. Bertoni

C. OPTICS

Program Director: L. Bergstein

D. QUANTUM ELECTRONICS

Program Director: W. T. Walter

E. SOLID STATE AND MATERIALS

Program Director: H. J. Juretschke

F. WAVE-MATTER INTERACTIONS

Program Director: N. Marcuvitz

G. ELECTRIC POWER ENGINEERING

Program Director: E. Levi

EXCITATION OF LARGE CONCAVE SURFACES

L.B. Felsen, A. Green and A. Hessel

While the excitation of, and propagation along, convex surfaces has been thoroughly explored in the technical literature, much less attention has been given to the corresponding problems for concave boundaries. Nevertheless, concave configurations are of importance in a variety of applications including ground wave propagation over terrain with smooth hills and depressions, scattering from reflectors and similar open structures, scattering from configurations backed by large cylindrical or spherical cavities, and mutual coupling in dome-shaped conformal arrays. Of special interest, and most difficult to analyze, are the fields observed on or near the surface when the source is also situated on or near the surface.

The most fundamental difference between the excitation of concave and convex surfaces or large (compared to the wavelength) radius of curvature is the absence in the former of a geometrical shadow region, from which the source is invisible. This circumstance gives rise not only to a more intricate geometric-optical field comprising multiply reflected rays (Fig. 1) but also, in an alternative guided wave description, to the presence of whispering gallery modes which cling to the surface (Fig. 2) and, in the absence of dissipation, experience no attenuation. Their counterparts on a convex surface, the creeping waves, lose energy by tangential shedding along the propagation path. The problem is complicated further by the fact that ray optics is incapable of providing the field solution since the caustics of multiply reflected rays pile up near the boundary and thus invalidate the geometric-optical field evaluation there. A field representation in terms of whispering gallery modes only (with inclusion of a continuous mode spectrum for some surfaces), while valid, is inconvenient for calculation for large separation of source and observation points on a large-radius surface since the number of modes required can be substantial. These problems do not arise on a convex surface where the distant field in the shadow region is represented compactly by the dominant creeping wave.

To gain a better physical as well as quantitative understanding of these aspects of wave propagation on a concave surface, intensive studies have been carried out on the simplest prototype configuration, the interior of a perfectly reflecting circular cylinder excited by an axial line source.[1,2,3] A peculiarity of the cylindrical geometry is the presence, in addition to the whispering gallery modes, of a continuous guided mode spectrum which arises because of spurious reflections from the radial coordinate origin. Elimination of these spurious contributions leads to an asymptotic field representation in terms of an integral which can be manipulated so as to exhibit ray-optical contributions, whispering gallery mode contributions, a mixture of these, or a

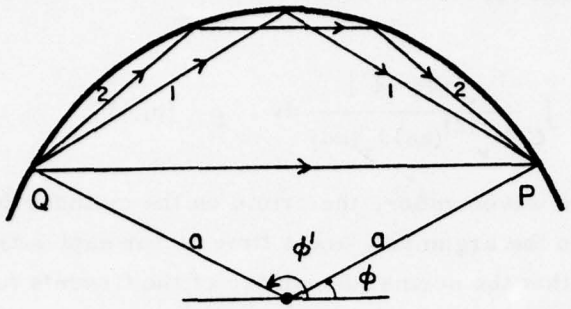


Fig. 1. Direct and multiply reflected rays.

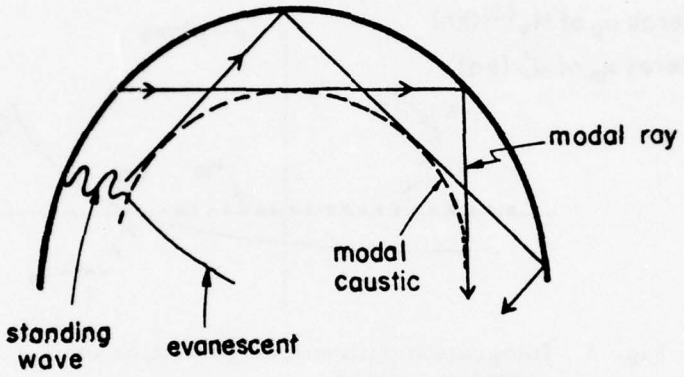


Fig. 2. Whispering gallery mode: modal field, modal rays and caustic.

formulation containing a reduced canonical integral analogous to the Fock integral for convex surfaces. The most effective choice depends on the parameters of the problem. The results so obtained can be generalized to apply to arbitrary concave surface shapes provided that the radius of curvature changes slowly over a wavelength interval. The validity of these variable-curvature solutions has been verified by comparison with exact calculations performed for a parabolic contour.[4]

We have extended the analysis to accommodate a surface impedance boundary condition. This is required when dealing with ground wave propagation and also with the concave array mutual coupling problem. Moreover, the results in the literature are inadequate for tracking the field to the vicinity of the source or, when source and observation points are fixed, to the limit of very large radius of curvature. For both cases, we have obtained a formulation in terms of the field for a flat boundary plus curvature dependent correction terms.

For the two-dimensional perfectly conducting circular cylinder, the relevant

propagation effects between source and observation points located on the boundary are contained in the integral

$$G_0(\underline{\rho}, \underline{\rho}') = \frac{1}{i(\pi ka)^2} \int_C \frac{e^{i\nu|\phi - \phi'|}}{H_\nu'^{(2)}(ka) J_\nu'(ka)} d\nu , \quad \underline{\rho} = (\rho, \phi) \tag{1}$$

where k is the free-space wavenumber, the prime on the cylinder functions denotes the derivative with respect to the argument, and a time factor $\exp(-i\omega t)$ is implied. It has been assumed in Eq. (1) that the normal derivative of the Green's function vanishes at $\rho = a$; this makes G_0 proportional to the axial component of magnetic field. The contour C and the singularities of the integrand in the complex ν-plane are shown in Figure 3.

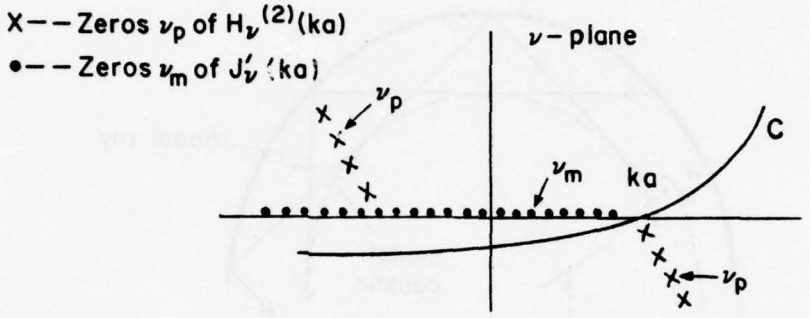


Fig. 3. Integration path and singularities in complex ν-plane.

Contributions from the pole singularities ν_m arising from

$$J_{\nu_m}'(ka) = 0 , \quad m = 1, 2, \cdots \tag{2}$$

are found to describe whispering gallery modes. While Eq. (2) has an infinite number of real solutions as indicated in Fig. 3, only those with $\mathrm{Re}\ \nu_m > 0$ represent spectral contributions in the angular transmission representation which includes also a continuous spectrum.

The alternative representations noted above are obtained from Eq. (1) by various analytical procedures including contour deformation, residue evaluation, saddle point techniques, traveling wave expansion of $1/J_\nu'(ka)$, partial summation of the resulting series, and the like. For most of these manipulations, it suffices to consider the simplified form of Eq. (1),

$$G_0 \sim -\frac{e^{iks}}{2\pi ka}\left(\frac{ka}{2}\right)^{2/3}\int_{\hat{C}}\frac{Ai(t)}{Ai'(t)}e^{i\gamma t}dt\ ,\quad \gamma = \left(\frac{ka}{2}\right)^{1/3}\frac{s}{a}\ , \tag{3}$$

where $s = a|\phi - \phi'|$ denotes the arc length between source and observation points, and the contour $\hat{C}$ is inferred from C by the mapping $\nu = ka + (ka/2)^{1/3}t$. In the near field or infinite plane limits (i.e., for small γ), Eq. (3) may be reduced to

$$G_0 \sim \frac{i}{2}H_0^{(1)}(ks)\left\{1 + \frac{e^{-i\pi/4}}{4}\sqrt{\pi}\,\gamma^{3/2} - \frac{i7}{60}\gamma^3 + O(\gamma^{9/2})\right\}\ , \tag{4}$$

where the factor outside the braces represents the Green's function for an infinite plane surface. When the curvature between fixed source and observation points changes from concave to convex (i.e., "a" is transformed continuously into (-a)), one may show that G_0 in Eq. (3) becomes

$$G_0 \sim \frac{e^{iks}}{2\pi ka}\left(\frac{ka}{2}\right)^{2/3}\int_{-\infty}^{\infty}\frac{w_1(t)}{w_1'(t)}e^{i\bar{\gamma}t}dt\ ,\quad \bar{\gamma} = \left(\frac{k|a|}{2}\right)^{1/3}\frac{s}{|a|}\ , \tag{5}$$

the known Green's function for a convex boundary. Here, $w_1 = Ai - iBi$.

Generalization to variable radius of curvature a(s) involves replacement of γ by $(k/2)\int_0^s [f(\mu)]^{-2}d\mu$, $f(s) = [ka(s)/2]^{1/3}$, and other related changes. To accommodate the impednace boundary condition, one generalizes the integral in Eq. (1) by employing $(H_\nu'^{(2)} - iZ\,H_\nu^{(2)})$ and $(J_\nu' - iZ\,J_\nu)$ instead of $H_\nu'^{(2)}$ and J_ν', respectively, with Z denoting the normalized surface impedance.

U.S. Army Research Office
DAHC04-75-G-0152

L.B. Felsen

REFERENCES

1. B. Ye. Kinber, "On the Diffraction of Electromagnetic Waves by a Concave Surface of a Circular Cylinder," Radio Tekh. Elektron., 6, No. 8, 1273-1283 (1961).
2. V.M. Babich and V.S. Buldyrev, "Asymptotic Methods of Short Wave Diffraction," (Moscow, USSR: Nauka, Chapter 11, Section 4).
3. W. Wasylkiwskyj, "Exact and Quasi-Optic Diffraction Within a Concave Cylinder," IEEE Trans. on Antennas and Propagation, AP-23, No. 4, 480-492 (July 1975).
4. V.M. Buldyrev and A.I. Lanin, "Asymptotic Formulas for a Wave Propagating Along a Concave Surface: Limits of Their Applicability," Radio Tekh. Elektron., 20 (1975).

RADIATION BY SOURCES IN A UNIDIRECTIONALLY CONDUCTING PLANE

A. Hessel, S. Siddiqi and J. Shmoys

A spiral antenna is a possibly useful phased array element. It is likely to have relatively high bandwidth and scan volume and can be made inexpensively. In order to study the performance of such an element the boundary value problem for a sheath model of the multi-arm spiral was formulated as a Galerkin-type procedure. In order to get an understanding of the convergence of the Galerkin procedure in this type of boundary value problem, we formulated a closely related two-dimensional problem in which the number of space harmonic modes is N (rather than N^2 for the three dimensional spiral structure). This reduces the computer time and permits the inclusion of a large number of space harmonics.

In a spiral element the circular source region is surrounded by a ring shaped sheath region in which conduction occurs along the wires only (cf. Figures 1 (a),(c)). The surface current in the source region is prescribed. The corresponding two-dimensional model has a strip source region and a strip sheath region on each side of the source, as shown on Figures 1(b),(d). In both cases the boundary value problem must be formulated so that for a given source current distribution the component of the electric field along the conduction direction must vanish in the x-y plane. In both cases it is simplest to follow the procedure outlined below:

(1) Expand the unknown current distribution in the conduction (sheath) region in a suitable set of functions and truncate the expansion. This introduces a finite number of coefficients to be determined.

(2) In terms of the unknown coefficients defined in (1), the magnetic field in the plane $z = 0^+$ is now completely defined; hence the electric field can be calculated.

(3) The component of the electric field in the conduction direction now must be set to zero. The essence of the Galerkin procedure is to project the function describing this field component onto the finite dimensional function space defined in (i) for the representation of the current distribution. This will yield as many equations as there are unknown coefficients.

(4) The set of equations is then solved and the physical quantities of interest calculated.

An alternative procedure can be used instead of (3) and (4) above: if the rms value of the tangential component of the electric field is calculated and then minimized with respect to all of the parameters used in the representation of the current distribution, the procedure is guaranteed to converge.

In this investigation we used three different representations of the current density distribution with the Galerkin method and a "quasi-static" representation in the variational method. For simplicity only broadside phasing will be considered, although the extension to other radiation directions is straightforward.

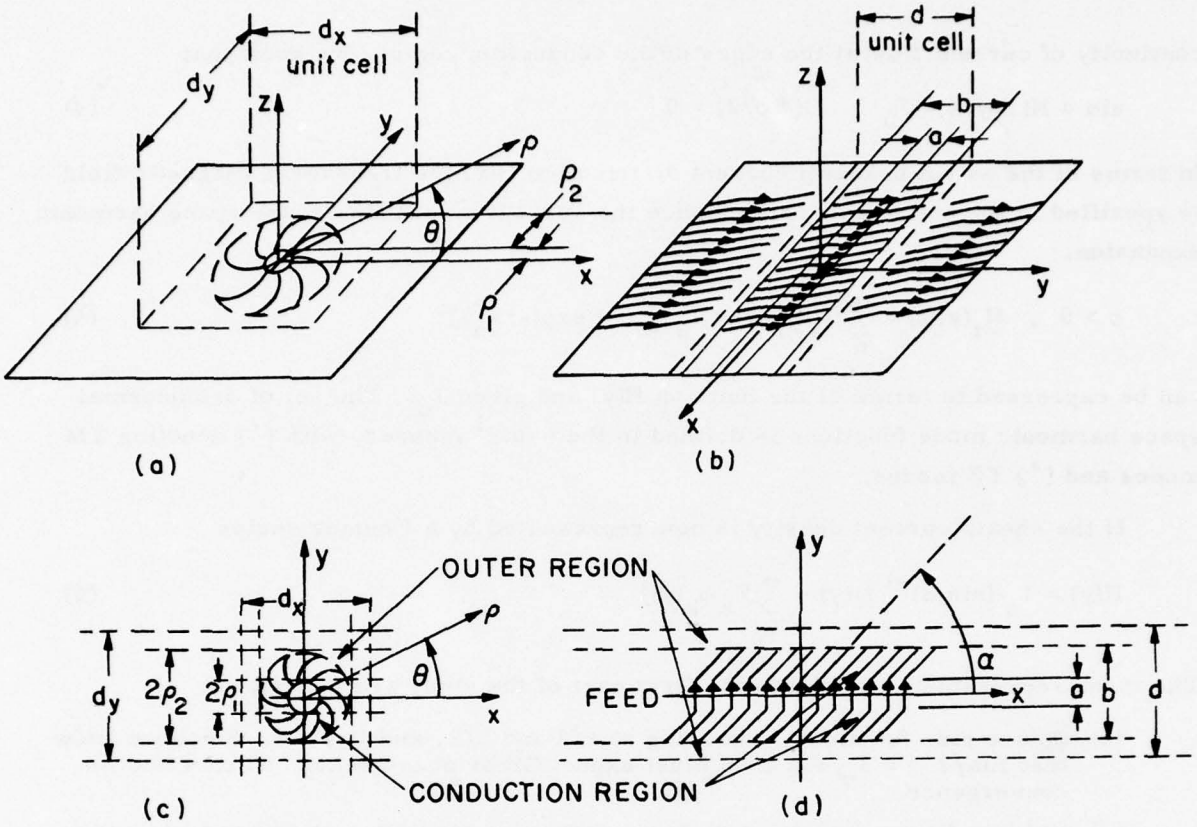


Fig. 1. Spiral and strip sheath elements.

A. Formulation of the Problem

Let us consider a unit cell as shown in Figs. 1 (b),(d), with the source surface current being given:

$$\underline{J}_s = 2J_o \underline{y}_o \quad , \quad |y| < a/2 \tag{1}$$

By symmetry, the transverse magnetic field in the z = 0+ plane is

$$\underline{H}_t = -\frac{1}{2}\underline{z}_o \times \underline{J}_s = J_o \underline{x}_o \quad , \quad |y| < a/2 \tag{2}$$

The thin wires in the conducting sheath region are characterized by perfect conductivity in the wire direction, $\underline{t}_o = \cos\alpha\, \underline{x}_o + \sin\alpha\, \underline{y}_o$, and zero conductivity in the perpendicular direction, so that the transverse magnetic field in the sheath region is

$$\underline{H}_t = -\underline{z}_o \times \underline{t}_o H(y) \quad , \quad a/2 < |y| < b/2 \tag{3}$$

where $2H_y$ is the complex current amplitude of the induced current. Finally, in the outer region, $b/2 < y < d/2$, the transverse magnetic field vanishes. From the

continuity of current flow at the edges of the conduction region, we know that

$$\sin \alpha \, H(\pm a/2) = J_o \quad ; \quad H(\pm b/2) = 0 \tag{4}$$

In terms of the as yet unknown current distribution H(y) the transverse magnetic field is specified in the entire unit cell. Hence the current amplitudes of the space harmonic expansion,

$$z > 0 \quad , \quad \underline{H}_t(y,z) = \sum_n (I'_n \underline{h}'_n(y) + I''_n \underline{h}''_n(t)) \exp(-j k_n z) \tag{5}$$

can be expressed in terms of the function H(y) and given J_o. The set of orthonormal space harmonic mode functions is defined in the usual[1] manner, with ($'$) denoting TM modes and ($''$) TE modes.

If the sheath current density is now represented by a Fourier series

$$H(y) = J_o \cdot (\sin \alpha)^{-1} \left[f(y) + \sum_k F_k \psi_k(y) \right] \tag{6}$$

The three representations used in the first part of the study are as follows:

(1) ψ_k are sine functions, vanishing at a/2 and b/2, and $f(y) \equiv 0$. Since we know that $H(a/2 -) = J_o / \sin \alpha$ we must expect Gibbs phenomenon and its effect on convergence.

(2) f(y) is a linear function satisfying continuity conditions at a/2 and b/2 and ψ_k vanish at a/2 and b/2.

(3) $f(y) \equiv 0$, and ψ_k vanish at b/2 but not at a/2: $\psi'_k = 0$ at a/2.

Since the space harmonic expansion is orthonormal, I'_n and I''_n can be readily calculated in terms of F_k's. The electric field can then be calculated as follows:

$$z > 0 \quad \underline{E}_t(y,z) = \sum_n (Z'_n I'_n \underline{e}'_n(y) + Z''_n I''_n \underline{e}''_n(y)) \exp(-j k_n z) \quad . \tag{7}$$

The remaining boundary condition, that of vanishing of the electric field along the wires,

$$\underline{E}_t(y,0) \cdot \underline{t}_o = 0 \quad a/2 < y < b/2 \quad , \tag{8}$$

can be imposed by setting the scalar products of the left-hand side of Eq. (8) with the sef of functions ψ_k be zero.

$$\int_{a/2}^{b/2} \underline{E}_t(y,0) \cdot \underline{t}_o \psi_k(y) \, dy = 0 \quad , \quad k = 1, \cdots, K \tag{9}$$

The use of the same truncated set of ψ_k in Eq. (6) and in enforcing Eq. (8) is the essence of the Galerkin procedure. The system of linear equations obtained is of order K. The

coefficients of this system involve summations over the space harmonics, which are also truncated. Thus we have

$$\sum_k A_{mk} F_k + B_m = 0 \qquad m = 1, 2, \cdots, K \tag{10}$$

where

$$A_{mk} = \sum_n \langle \psi_m | \underline{e}_n \cdot \underline{t}_o \rangle Z_n \langle \underline{e}_n \cdot \underline{t}_o | \psi_k \rangle \quad , \tag{11}$$

$$B_m = \sum_n Z_n (I_{n\,FEED} + \langle \underline{e}_n \cdot \underline{t}_o | f(y) \rangle) \langle \psi_m | \underline{e}_n \cdot \underline{t}_o \rangle \quad , \tag{12}$$

the summation extends over all, TE and TM, space harmonics. The scalar product notation is defined by

$$\langle \Phi | \psi \rangle = \int_{a/2}^{b/2} \Phi^*(y)\, \psi(y)\, dy$$

and $I_{n\,FEED}$ denotes a similar scalar product of the n-th space harmonic with the impressed magnetic field in the feed (center) region. Once the coefficients F_k are determined, all the physically relevant quantities can be readily calculated.

In the variational method[2] we made use of the function

$$\begin{aligned} f(y) = {} & U\left((y - a/2)^{1/2} - ((b-a)/2)^{1/2} - (y - b/2)(2/b - a))^{1/2}\right) \\ & + V\left((b/2 - y)^{1/2} + (y - b/2)(2/(b-a))^{1/2}\right) - (2y - b)/(b - a) \end{aligned} \tag{13}$$

in the representation of the magnetic field, Eq. (6), with ψ_k being sine functions. This expression contains $2 + K$ complex constants or $4 + 2K$ real parameters. The mean square tangential field is then given by

$$\int_{a/2}^{b/2} |\underline{E}_t \cdot \underline{t}_o|^2 dy = Q(U_r, U_i, V_r, V_i, F_{kr}, F_{ki}) \quad , \tag{14}$$

a quadratic form in the parameters used in Eq. (13) and real and imaginary parts of the Fourier coefficients.

The coefficients of this quadratic form are sums over space harmonics. They are obtained by putting Eq. (13) into Eq. (6), then calculating $\underline{E}_t$ using Eqs. (5) and (7), finally putting the result in Eq. (14) and collecting all terms in U_r^2, $U_r U_i$, U_i^2, etc. Since the tangential electric field in the sheath region should be zero, we get the best approximation by minimizing the quadratic form. The convergence of this procedure is self-evident.[2] The choice of the function f(y) in Eq. (13) was dictated by the fact that it satisfies the continuity conditions and has the correct quasi-static square root behavior at the ends of the interval. This procedure was carried out for the case $K = 1$ only.

B. Discussion of Results

One way in which we can examine the convergence of the method is by comparing the magnetic field distribution calculated using different numbers of Fourier series terms (K) and different numbers of space harmonics (from -N to N). In order to avoid relative convergence difficulties the ratio of N : K was kept at 2 which provides approximately the same spacial resolution in both functions. The approximate magnetic field distribution can be calculated in two ways: (1) directly from the Fourier series in the conduction region (with J_o and 0 elsewhere) and (2) from space harmonic series in the entire unit cell. One indication of accuracy is the agreement of the latter results with the former, particularly in the source and outer regions where the field is prescribed. Such a comparison, for N = 20, K = 10, using sine series, representation (a), is shown on Figure 2. Conduction direction in this and succeeding cases (except for the last one) is at right angles to the elements, $\alpha = 90^{\circ}$, so that it makes no difference whether the sheath region is unidirectionally or omnidirectionally conducting. We see that the agreement between the two curves is quite satisfactory. The same comparison for K = 50, N = 100 (Fig. 3), is even better. Figure 4 shows a comparison of Fourier series sum for representations (2), sine series with linear transition term and (3), cosine series. Both are a clear improvement over the sine series results.

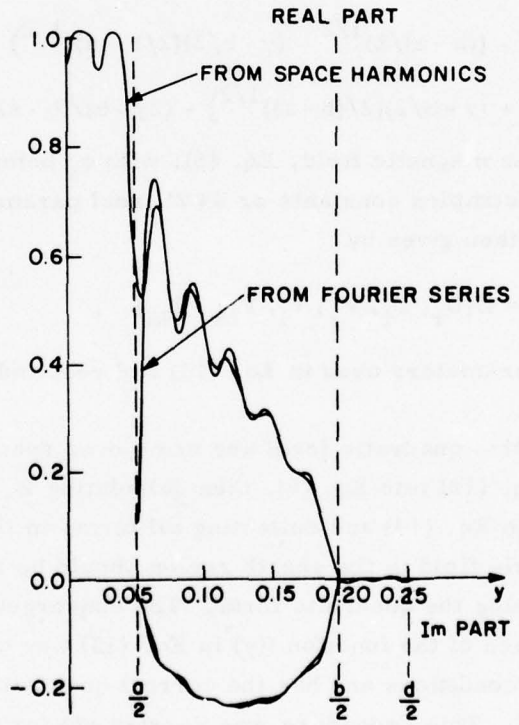


Fig. 2. Normalized magnetic field distribution. N = 20 K = 10 representation (1).

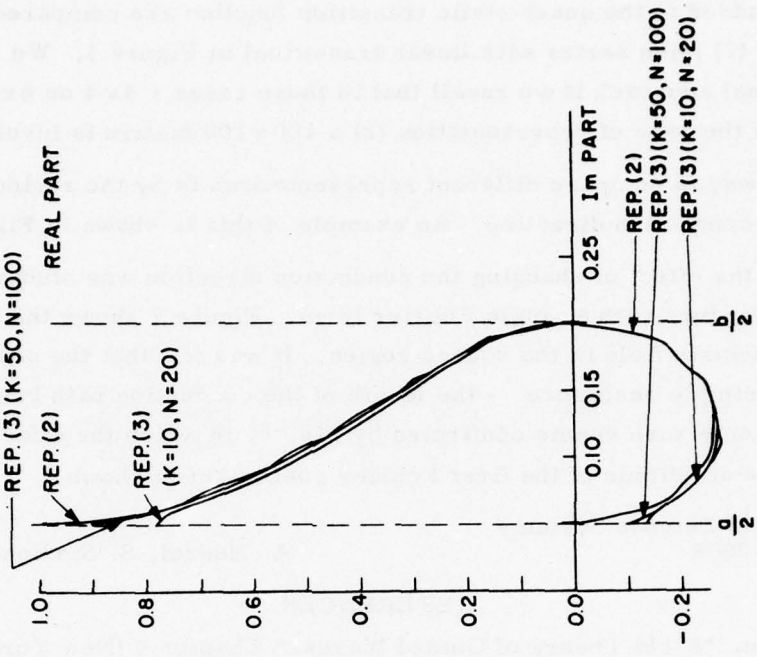


Fig. 4. Normalized magnetic field distribution.
Representation (2) and (3).

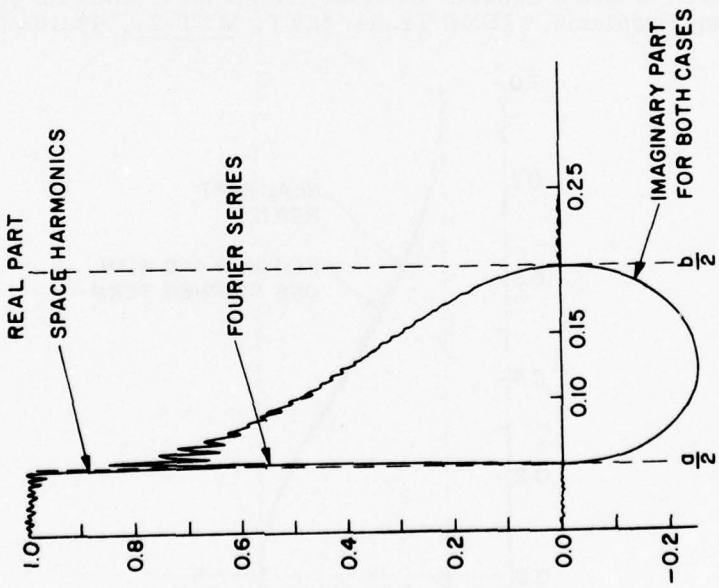


Fig. 3. Normalized magnetic field distribution.
K = 50 N = 100 representation (1).

The results of the variational approach without any Fourier terms, and with one Fourier term added to the quasi-static transition function are compared with those for representation (2) (sine series with linear transition) in Figure 5. We see the advantage of the variational approach if we recall that in these cases a 4x4 or 6x6 matrix is inverted while in the case of representation (2) a 100x100 matrix is involved.

Another way to compare different representations is by the residual electric field parallel to the conduction direction. An example of this is shown in Figure 6.

Finally, the effect of changing the conduction direction was studied, using the variational procedure with a single Fourier term. Figure 7 shows the effect of tilt angle α on induced electric field in the source region. It was felt that the peak shown at 20° is related to simple resonance -- the length of the conduction path being a half wavelength. This conjecture seems confirmed by Fig. 8, in which the effect of changing the tilt angle on the amplitude of the first Fourier coefficient is shown.

Ballistic Missile Defense Systems
DASG60-76-C-0006

A. Hessel, S. Siddiqui and J. Shmoys

REFERENCES

1. R.E. Collin, "Field Theory of Guided Waves," Chapter 9 (New York: McGraw Hill, 1960).
2. J.B. Davies, "A Least Squares Boundary Residual Method for the Numerical Solution of Scattering Problems," IEEE Trans. MTT, MTT-21, 99-103 (February 1973).

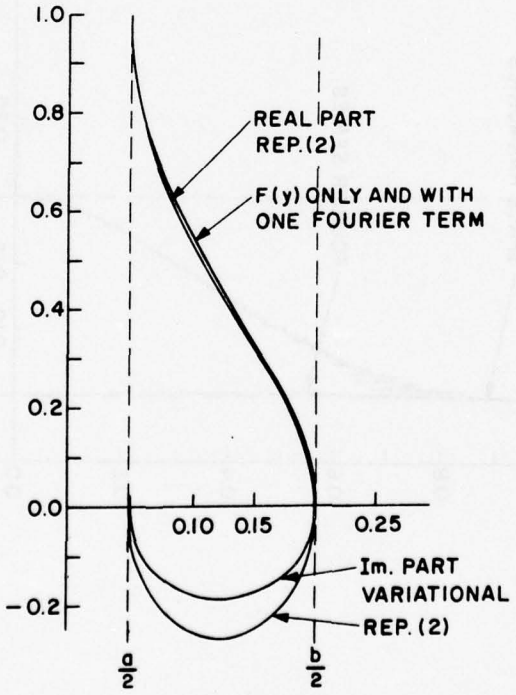


Fig. 5. Normalized magnetic field distribution. Variational results.

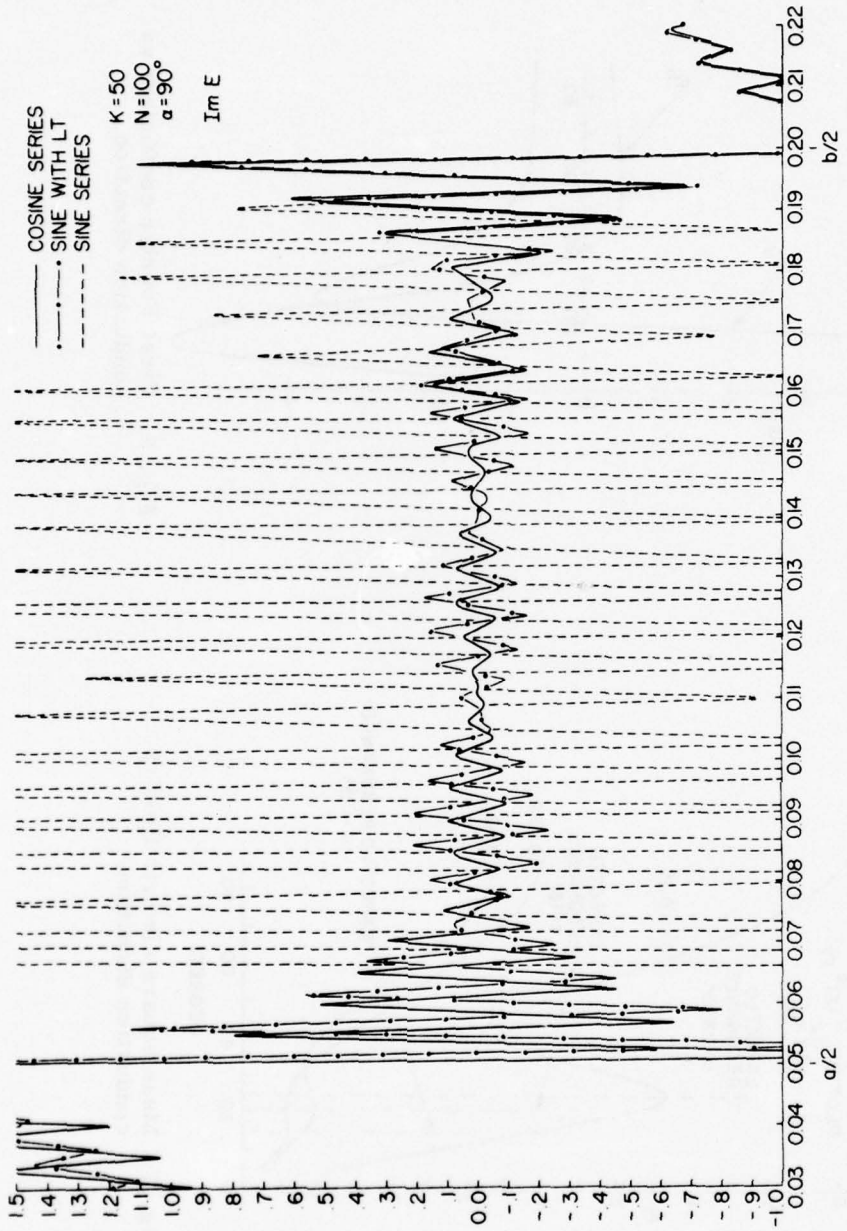


Fig. 6. Tangential electric field distribution (quadratic component).

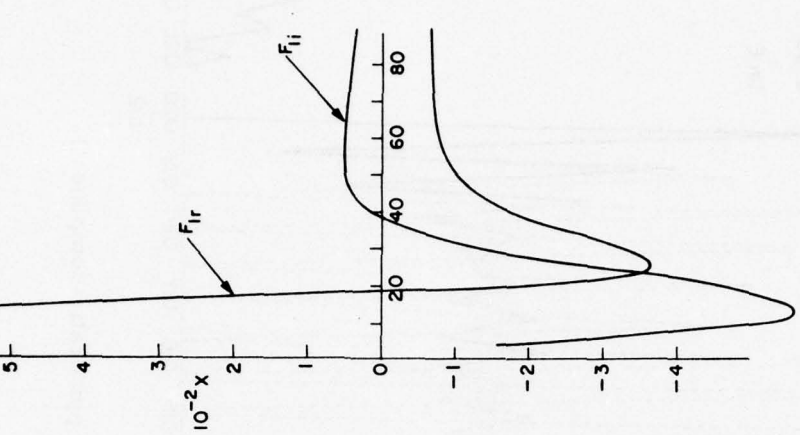


Fig. 8. First Fourier coefficient vs. conduction direction.

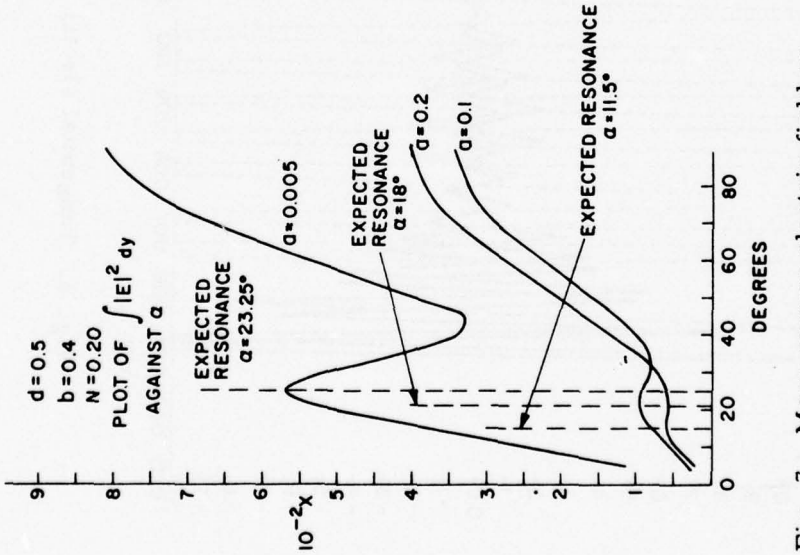


Fig. 7. Mean square electric field vs. conduction direction.

TRANSIENT DIFFRACTION BY A SEMI-INFINITE CONE

K.K. Chan and L.B. Felsen

The time-dependent response due to a point source in the presence of a semi-infinite acoustically soft or hard cone can be found through Laplace inversion of the time harmonic Green's functions.[1] The solution obtained in Ref. 1, while generally valid, is given as an infinite series which is useful for computation only in those parameter ranges where it converges rapidly: at long observation times. In the following, a different approach is used; we employ a "quasi-optic" integral representation of the time harmonic Green's function, which leads to a new form of the time dependent solution. The result, which contains an integral, can be reduced further if we restrict ourselves to the evaluation of the scattered field by a small-angle cone, with the source located on the cone axis. As will be shown, solutions for all observation times are obtained in remarkably simple closed forms involving only elementary functions provided that the observation point lies in a region that excludes the rays reflected from the cone surface according to the laws of geometrical optics. This region of applicability accommodates important applications, in particular, the on-axis back-scattered fields. To confirm the validity of the closed form results, the early-time and long-time behavior of the transient field are evaluated as special cases and compared, respectively, with solutions obtained independently by different approaches. Both the scalar Dirichlet and Neumann problems, and the vector dipole problems (leading to the dyadic Green's functions), can be treated in this manner.

A. Soft Cone (Dirichlet Condition)

1. Formulation and Solution

The time dependent Green's function $\hat{G}(\underline{r}, \underline{r}'; t, t')$ can be obtained via Laplace inversion from the time-harmonic Green's function (with time dependence $e^{-i\omega t}$ suppressed throughout), which satisfies the three-dimensional wave equation

$$(\nabla^2 + k^2)\, G(\underline{r}, \underline{r}') = -\delta(\underline{r} - \underline{r}') \tag{1}$$

and boundary conditions

$$\text{(a)}\quad G(\underline{r}, \underline{r}') = 0 \quad \text{at} \quad \theta_0 \tag{1a}$$

$$\text{(b)}\quad \text{radiation condition at } r \to \infty \tag{1b}$$

$$\text{(c)}\quad \text{tip condition } (G(\underline{r}, \underline{r}') \text{ finite}) \text{ at } r = 0 \tag{1c}$$

The geometry of the configuration and the assumed spherical coordinate system are shown in Figure 1.

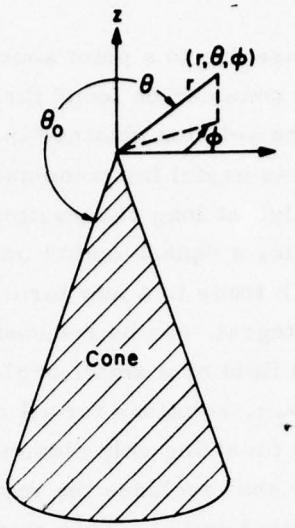


Fig. 1. Physical configuration.

The solution for $G(\underset{\sim}{r}, \underset{\sim}{r}')$ can be written as the sum of a free-space field $G_0(\underset{\sim}{r}, \underset{\sim}{r}')$ and a scattered field $G_s(\underset{\sim}{r}, \underset{\sim}{r}')$, due to the presence of the cone, as follows:[2]

$$G(\underset{\sim}{r}, \underset{\sim}{r}') = G_0(\underset{\sim}{r}, \underset{\sim}{r}') + G_s(\underset{\sim}{r}, \underset{\sim}{r}') \tag{2}$$

where

$$G_0(\underset{\sim}{r}, \underset{\sim}{r}') = \frac{e^{ik|\underset{\sim}{r}-\underset{\sim}{r}'|}}{4\pi|\underset{\sim}{r}-\underset{\sim}{r}'|} \tag{3a}$$

$$G_s(\underset{\sim}{r}, \underset{\sim}{r}') = \frac{1}{8\pi krr'} \sum_{m=0}^{\infty} \varepsilon_m \cos m(\phi-\phi') \cdot$$

$$\cdot \int_{-1/2-i\infty}^{-1/2+i\infty} d\nu\,(2\nu+1)\, h_\nu^{(1)}(kr)\, h_\nu^{(1)}(kr')\, \frac{\Gamma(\nu+m+1)}{\Gamma(\nu-m+1)} \cdot$$

$$\cdot \frac{P_\nu^{-m}(\cos\theta)\, P_\nu^{-m}(\cos\theta')\, P_\nu^{-m}(-\cos\theta_0)}{\sin[(\nu-m)\pi]\, P_\nu^{-m}(\cos\theta_0)} \tag{3b}$$

with

$$\varepsilon_0 = \frac{1}{2}; \quad \varepsilon_m = 1, \quad m \neq 0 .$$

Here, $P_\nu^{-m}(x)$ is the Legendre function of order ν, degree (-m) and argument x, while

$z_\nu(x)$ is a spherical Bessel function

$$z_\nu(x) = \sqrt{\frac{\pi x}{2}}\, Z_{\nu+1/2}(x) \quad , \tag{3c}$$

with $Z_\nu(x)$ denoting a solution of the conventional Bessel equation (in Eq. (3b), $z_\nu \equiv h_\nu^{(1)}$). To get the result in Eq. (3b), the condition $(\theta+\theta') < (2\theta_0 - \pi)$ has to be imposed (Reference 3). This restriction on $(\theta+\theta')$ excludes from the range of applicability that domain which contains the rays reflected from the sides of the cone according to the laws of geometrical optics (see Figure 2).

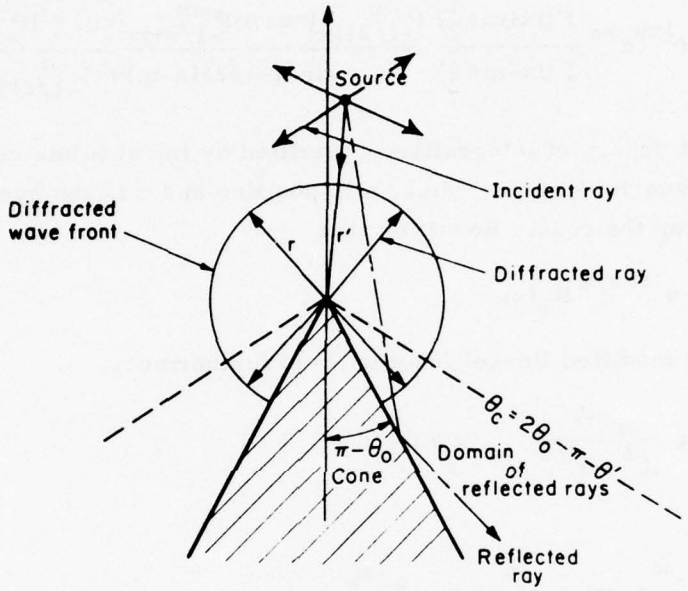


Fig. 2. Ray-optical domains.

To convert the representation in Eq. (3b) into a form that permits direct Laplace inversion, we first employ the change of variable

$$\nu = -\frac{1}{2} + ix \tag{4}$$

(The Laplace inversion of G_0 in Eq. (3a) yields

$$\hat{G}_0(\underline{r}, \underline{r}'; t, t') = [4\pi\,|\underline{r}-\underline{r}'|\,]^{-1}\,\delta\left(t - t' - \frac{|\underline{r}-\underline{r}'|}{c}\right).)$$

Introducing the Hankel function product formula[4]

$$H_{ix}^{(1)}(kr)\, H_{ix}^{(1)}(kr') = \frac{1}{\pi} \int_{i\infty}^{-i\infty} dw\, H_0^{(1)}\left(k\sqrt{r^2 + r'^2 + 2rr'\cos w}\right) e^{-x(w-\pi)} \tag{5}$$

and letting $y = iw$, one transforms Eq. (3b) into

$$G_s(\underset{\sim}{r}, \underset{\sim}{r}') = \int_{-\infty}^{\infty} dy\, H_0^{(1)}\left(k\sqrt{r^2 + r'^2 + 2rr'\cosh y}\right) A(y) \tag{6}$$

where $A(y) \equiv A(\theta, \theta', \phi, \phi'; y)$ is given by

$$A(y) = \frac{i}{8\pi\sqrt{rr'}} \sum_{m=0}^{\infty} \varepsilon_m \cos m(\phi - \phi') \cdot \int_{-\infty}^{\infty} dx\, x e^{ixy} e^{x\pi} \frac{\Gamma(ix+m+\frac{1}{2})}{\Gamma(ix-m+\frac{1}{2})} \frac{P^{-m}_{-1/2+ix}(\cos\theta) P^{-m}_{-1/2+ix}(\cos\theta') P^{-m}_{-1/2+ix}(-\cos\theta_0)}{\sin[(-1/2+ix-m)\pi]\, P^{-m}_{-1/2+ix}(\cos\theta_0)} \tag{6a}$$

The interchange of orders of integration is justified by the absolute convergence of the double integral. Now let $k = is/c$, where s is positive and c is the speed of light in the medium surrounding the cone. Recalling that

$$H_\nu^{(1)}(iz) = \frac{2}{\pi i} e^{-i\nu\pi/2} K_\nu(z) \tag{7}$$

where $K_\nu(z)$ is the modified Bessel function, and furthermore

$$K_0(x) = \int_1^{\infty} d\zeta \frac{e^{-x\zeta}}{(\zeta^2 - 1)^{1/2}} \quad , \quad x > 0 \tag{8}$$

we obtain

$$G_s(\underset{\sim}{r}, \underset{\sim}{r}') = \int_0^{\infty} dy\, B(y) \int_{f/c}^{\infty} d\xi \frac{e^{-s\xi}}{\sqrt{\xi^2 - f^2/c^2}} \tag{9}$$

where

$$B(y) = \frac{2}{\pi i}\left[A(y) + A(-y)\right] \tag{9a}$$

$$f \equiv f(y) = \sqrt{r^2 + r'^2 + 2rr'\cosh y} \tag{9b}$$

Interchanging the orders of integration, one finds after simply manipulations,

$$G_s(\underset{\sim}{r}, \underset{\sim}{r}') = \int_0^{\infty} d\xi\, e^{-s\xi}\, Q(\xi) \tag{10}$$

where

$$Q(\xi) = 0 \quad , \quad c\xi < r + r' \tag{10a}$$

$$= \int_0^{a(\xi)} dy \, \frac{B(y)}{\sqrt{\xi^2 - f^2/c^2}} \quad , \quad c\xi > r + r' \tag{10b}$$

with

$$a(\xi) = \cosh^{-1}\left[\frac{c^2\xi^2 - r^2 - r'^2}{2rr'}\right] \tag{10c}$$

In view of Eq. (10), which is itself in the form of a Laplace transform, the inversion of G_s into the time domain is trivial. It follows that

$$\hat{G}_s(\underset{\sim}{r}, \underset{\sim}{r}'; t, t') = \int_0^a dy \, \frac{B(y)}{\sqrt{\tau^2 - f^2/c^2}} \cdot U(\tau - \frac{r + r'}{c}) \tag{11}$$

where $U(x)$ is the Heaviside unit function, $a \equiv a(\tau)$ and $\tau = t - t'$. This solution can be further simplified. Substituting Eqs. (9a) and (6a) into Eq. (11) and performing an interchange of the order of integration, one may carry out the integration in y by utilizing the integral representation of the Legendre function[5]

$$P_{-1/2+ix}(\cosh a) = \frac{\sqrt{2}}{\pi} \int_0^a dy \, \frac{\cos xy}{\sqrt{\cosh a - \cosh y}} \tag{12}$$

Thus, from Eq. (12) and properties of the gamma function, we obtain

$$\hat{G}_s(\underset{\sim}{r}, \underset{\sim}{r}'; t, t') = \int_0^\infty dx \, \hat{F}(x) \, P_{-1/2+ix}(\cosh a) \, U(\tau - \frac{r + r'}{c}) \tag{13}$$

where

$$\cosh a = \frac{c^2\tau^2 - r^2 - r'^2}{2rr'} \quad , \tag{14a}$$

$$\hat{F}(x) = -\frac{c}{2rr'} \sum_{m=0}^{\infty} \epsilon_m \cos m(\phi - \phi') \, \frac{x \tanh \pi x}{\cosh \pi x}$$

$$\cdot \frac{P^{-m}_{-1/2+ix}(\cos\theta) \, P^{-m}_{-1/2+ix}(\cos\theta')}{\Gamma(ix - m + 1/2) \, \Gamma(-ix - m + 1/2)} \, \frac{P^{-m}_{-1/2+ix}(-\cos\theta_0)}{P^{-m}_{-1/2+ix}(\cos\theta_0)} \tag{14b}$$

The expression in Eq. (13) yields the time-dependent tip diffracted field for arbitrary source and observation points provided, however, that $\theta + \theta' < 2\theta_0 - \pi$. The solution describes a spherical pulse centered at the cone tip and arriving at the

observation point at $\tau = (r + r')/c$, i.e., after the time interval required for the incident pulse to travel from the source point to the tip and for the diffracted pulse to travel from the tip to the observation point (Figure 2).

2. Closed Form for On-Axis Source and Small Cone Angle

To simplify the general result, we assume first that the source point is located on the cone axis, i.e., $\theta' = 0$. Then Eq. (13) reduces to

$$\hat{G}_s(\underset{\sim}{r}, \underset{\sim}{r}'; t, t') = \frac{-c}{4\pi rr'} \int_0^\infty dx\, x \tanh \pi x\, P_{-1/2+ix}(\cos\theta)\, P_{-1/2+ix}(\cosh a) \cdot \frac{P_{-1/2+ix}(-\cos\theta_0)}{P_{-1/2+ix}(\cos\theta_0)}\, U(\tau - \frac{r+r'}{c}) \tag{15}$$

Furthermore, we specialize to small cone angles $\theta_0 \approx \pi$, for which the following approximation applies:[6]

$$\frac{P_{-1/2+ix}(-\cos\theta_0)}{P_{-1/2+ix}(\cos\theta_0)} \sim -\frac{\pi}{\cosh \pi x \ln[(\frac{\pi - \theta_0}{2})^2]}, \quad \theta_0 \approx \pi \tag{16}$$

Eq. (15) is then reduced to:

$$\hat{G}_s(r, r'; t, t') \sim \frac{c}{4rr' \ln[(\frac{\pi - \theta_0}{2})^2]} \cdot \int_0^\infty dx\, \frac{x \tanh \pi x}{\cosh \pi x} P_{-1/2+ix}(\cos\theta) P_{-1/2+ix}(\cosh a)\, U(\tau - \frac{r+r'}{c}) \tag{17}$$

The integrand in Eq. (17) is a Mehler transform. By employing the formula[7]

$$\int_0^\infty dx\, \frac{x \tanh \pi x}{\cosh \pi x} P_{-1/2+ix}(w) P_{-1/2+ix}(z) = \frac{1}{\pi(w+z)}, \tag{17a}$$

one obtains in closed form,

$$\hat{G}_s(\underset{\sim}{r}, \underset{\sim}{r}'; t, t') \sim \frac{c}{4\pi rr' \ln[(\frac{\pi - \theta_0}{2})^2]} \frac{1}{\cosh a + \cos\theta}\, U(\tau - \frac{r+r'}{c}) \tag{18}$$

or, written explicitly,

$$\hat{G}_s(\underset{\sim}{r}, \underset{\sim}{r}'; t, t') \sim \frac{c}{8\pi rr' \ln[(\frac{\pi - \theta_0}{2})^2]} \frac{1}{b^2 + d^2}\, U(\tau - \frac{r+r'}{c}) \tag{19}$$

where

$$b^2 = \frac{c^2\tau^2 - (r+r')^2}{4rr'}, \quad d = \cos\frac{\theta}{2} \tag{19a}$$

This expression is remarkably simple and quite unexpectedly so; however, it is subject to the restrictions $\theta < 2\theta_0 - \pi$, $\theta_0 \approx \pi$, mentioned earlier. It can be shown that the solution agrees with independently derived results for early-time and long-time responses, thereby confirming its validity in these limiting cases.

Analogous simple results can be developed for the Neumann (acoustically hard) boundary condition, and for the electromagnetic vector dipole field in the presence of a perfectly conducting cone.

Army Research Office
DAHC04-75-G-0152

L.B. Felsen

REFERENCES

1. F. Oberhettinger and R.F. Dressler, "Electromagnetic Fields in the Presence of Ideally Conducting Conical Structures," J. Appl. Math. and Phys. (ZAMP), 2, 937-950 (1971).
2. L.B. Felsen and N. Marcuvitz, "Radiation and Scattering of Waves," (Englewood Cliffs, N.J.: Prentice Hall, Inc., 1973), Sec. 6.8, 703.
3. L.B. Felsen, "Plane Wave Scattering by Small Angle Cones," IRE Trans. on Antennas and Propagation, AP-5, 404 (1957).
4. ibid., Ref. 2, Appendix 6B, 718.
5. W. Magnus and F. Oberhettinger, "Formulas and Theorems for the Special Functions of Mathematical Physics," New York: Springer Verlag, 1966), 184.
6. ibid., Ref. 2, 708.
7. H. Bateman, et.al., "Higher Transcendental Functions," (New York: McGraw-Hill, 1953), 174.

ANALYSIS OF SURFACE WAVEGUIDE DISCONTINUITIES

E-W. Hu and L. Bergstein

Using a closed waveguide approach, we analyzed recently[1] the scattering of an even TE surface wave mode impinging upon a step discontinuity of an open dielectric slab waveguide. We report here the results for the same problems but with an even TM incident surface wave.

Marcuse[2] has solved the scattering problem using two different approximations. For the TE case, both approximations yield the same answer which agrees with our results. For the TM case, however, the results of the two approximations begin to diverge as the waveguide thickness increases beyond a quarter of a wavelength, with one of the approximations yielding negative radiation losses for large waveguide thicknesses. Our approach yields correct results irrespective of the waveguide thickness. Moreover, it is straightforward and applicable to multimode and other, more complex waveguide structures. It has the additional advantage that it yields information for the radiation pattern of the scattered waves.

A. Formulation of the Problem

The open dielectric slab waveguide with a discontinuity at $z = 0$ is shown in Figure 1(a). The thickness of the waveguide is $2d_1$ to the left of the discontinuity and $2d_2$ to the right. The dielectric constants in the various waveguide regions are given by

$$\varepsilon_L(x) = \begin{cases} \varepsilon_{r1}\varepsilon_0, & -d_1 < x < +d_1 \\ \varepsilon_{r2}\varepsilon_0, & |x| > d_1 \end{cases} \tag{1a}$$

for the left waveguide ($z < 0$), and

$$\varepsilon_R(x) = \begin{cases} \varepsilon_{r3}\varepsilon_0, & -d_2 < x < +d_2 \\ \varepsilon_{r4}\varepsilon_0, & |x| > d_2 \end{cases} \tag{1b}$$

for the right waveguide ($z > 0$), where ε_{ri} ($i = 1, 2, 3, 4$) is the relative dielectric permittivity and ε_0 is the permittivity of free space. It is assumed that $\varepsilon_{r1} > \varepsilon_{r2}$ and $\varepsilon_{r3} > \varepsilon_{r4}$.

We approximate the open waveguide by the corresponding closed structure shown in Fig. 1(b) and choose the distance a_1 of the perfectly conducting enclosure sufficiently large so that a further increase will have no discernible effect on the results.

Assuming that there is no field variation in the y direction, i.e., that $\partial/\partial y = 0$, we can decompose the fields into TE and TM modes. Moreover, from symmetry considerations it is clear that instead of solving the waveguide structure of Fig. 1(b) we can deal with the simpler geometries shown in Figs. 2(a) and 2(b) for odd and even TM modes, respectively.

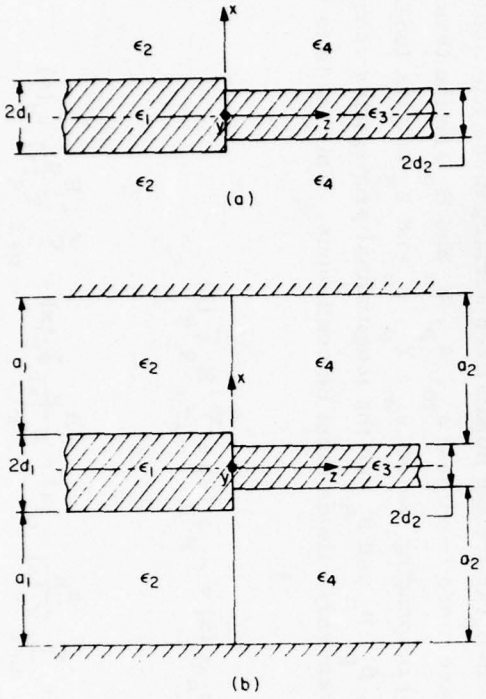


Fig. 1. (a) Planar surface waveguide with step discontinuity in geometry and in dielectric constant; (b) the corresponding closed waveguide structure.

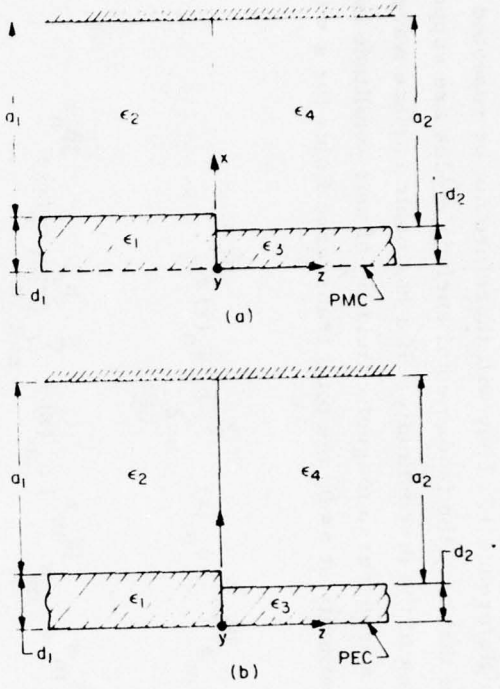


Fig. 2. The reduced dielectric slab surface of Fig. 1(b): (a) shows the reduced waveguide for the odd TM case with a perfect magnetically conducting bisecting wall; (b) shows the reduced waveguide for the even TM case with a perfect electrically conducting bisecting wall.

The mode functions and dispersion equations for the various waveguide regions are discussed and listed in Reference 1. They will therefore not be repeated here. For simplicity, we assume that only the fundamental surface modes are supported by the waveguides on both sides of the discontinuity. If a harmonic surface wave with a time variation $\exp(j\omega t)$ and a transverse magnetic field H_y of unit amplitude is incident from $-\infty$ towards the discontinuity at $z = 0$, the total transverse fields for $z < 0$ after scattering are given by

$$H_y(x,z) = (e^{-j\beta_m z} - a_m e^{j\beta_m z})\,\phi_m(x) - \sum_{n=2}^{\infty} b_n\,\psi_n(x)\,e^{j\beta_n z} \tag{2a}$$

and

$$E_x(x,z) = \frac{1}{y_m(x)}(e^{-j\beta_m z} + a_m e^{j\beta_m z})\,\phi_m(x) + \sum_{n=2}^{\infty} \frac{b_n}{y_n(x)}\,\psi_n(x)\,e^{j\beta_n z} \tag{3a}$$

Similarly, the total transverse fields for $z > 0$ are given by

$$H_y(x,z) = A_p e^{-j\beta_p z}\,\bar{\Phi}_p(x) + \sum_{q=2}^{\infty} B_q\,\bar{\psi}_q(x)\,e^{-j\beta_q z} \tag{2b}$$

and

$$E_x(x,z) = \frac{A_p}{Y_p(x)}\,e^{-j\beta_p z}\,\bar{\Phi}_p(x) + \sum_{q=2}^{\infty} \frac{B_q}{Y_q(x)}\,\bar{\psi}_q(x)\,e^{-j\beta_q z} \tag{3b}$$

In these equations, $\phi_m(x)$ and $\bar{\phi}_p(x)$ are the normalized surface mode functions, $\psi_n(x)$ and $\bar{\psi}_q(x)$ are the non-surface mode functions; a_m, A_p, b_n and B_q are the transverse field amplitudes for the corresponding waves; y_m, Y_p, y_n and Y_q are the longitudinal wave impedances; and β_m, β_p, β_n and β_q are the longitudinal propagation constants. At the junction $z = 0$, the transverse fields must be continuous. This leads to the following two sets of equations:

$$(1 - a_m)\,\phi_m(x) - \sum_{n=2}^{\infty} b_n\,\psi_n(x) = A_p\,\bar{\phi}_p(x) + \sum_{q=2}^{\infty} B_q\,\bar{\psi}_q(x) \tag{4}$$

and

$$\frac{1}{y_m(x)}(1 + a_m)\,\phi_m(x) + \sum_{n=2}^{\infty} \frac{b_n}{y_n(x)}\,\psi_n(x) = \frac{A_p}{Y_p(x)}\,\bar{\phi}_p(x) + \sum_{q=2}^{\infty} \frac{B_q}{Y_q(x)}\,\bar{\psi}_q(x) \tag{5}$$

Using the orthogonality properties of mode functions, Eqs. (4) and (5) reduce to

$$\left\{ \langle \phi_m, \frac{1}{\varepsilon_R} \bar{\phi}_p \rangle + \frac{\beta_m}{\beta_p} \langle \phi_m, \frac{1}{\varepsilon_L} \bar{\phi}_p \rangle + \langle \phi_m, \frac{1}{\varepsilon_R} \psi_q \rangle + \frac{\beta_m}{\beta_q} \langle \phi_m, \frac{1}{\varepsilon_L} \psi_q \rangle \right\} a_m$$

$$+ \sum_{n=2}^{\infty} \left\{ \langle \psi_n, \frac{1}{\varepsilon_R} \bar{\phi}_p \rangle + \frac{\beta_n}{\beta_p} \langle \psi_n, \frac{1}{\varepsilon_L} \bar{\phi}_p \rangle + \langle \psi_n, \frac{1}{\varepsilon_R} \bar{\psi}_q \rangle + \frac{\beta_n}{\beta_q} \langle \psi_n, \frac{1}{\varepsilon_L} \bar{\psi}_q \rangle \right\} b_n$$

$$= \langle \phi_m, \frac{1}{\varepsilon_R} \bar{\phi}_p \rangle - \frac{\beta_m}{\beta_p} \langle \phi_m, \frac{1}{\varepsilon_L} \bar{\phi}_p \rangle + \langle \phi_m, \frac{1}{\varepsilon_R} \bar{\psi}_q \rangle - \frac{\beta_m}{\beta_q} \langle \phi_m, \frac{1}{\varepsilon_L} \bar{\psi}_q \rangle \qquad (6)$$

with $m = p = 1$ and $q = 2, 3, 4, \ldots, \infty$. We solve this infinite system of linear equations for the amplitudes a_m and b_n by truncating the higher evanescent modes and solving the resulting finite set of linear equations numerically. The truncation process only results in a negligibly small error.

B. Numerical Results

We solved numerically the discontinuity problem (i.e., Eqs. (6)) for the same set of parameters as those reported by Marcuse.[2] Thus, we assumed that the refractive indices $n_i = \sqrt{\varepsilon_{ri}}$ of the waveguide media are $n_2 = n_4 = 1$ and $n_1 = n_3 = 1.432$ and have determined the scattered fields as a function of the normalized wave thickness $kd_1 = 2\pi d_1/\lambda$ with the thickness ratio d_2/d_1 as a parameter. The results for the radiation losses are shown in Fig. 3 for the case when $d_2/d_1 = 0.5$. The two dashed curves show

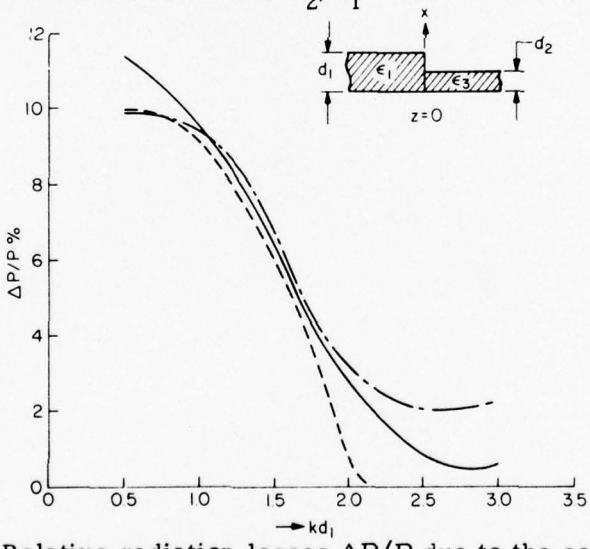


Fig. 3. Relative radiation losses $\Delta P/P$ due to the scattering of surface waves by the step discontinuity for the even TM case with $n_1 = n_3 = 1.432$, $n_2 = n_4 = 1$, and $d_2/d_1 = 0.5$. The losses are shown as a function of the normalized waveguide thickness $kd_1 = 2\pi d_1/\lambda$. Solid curve shows the result obtained in this work, the two dashed curves represent the results reported by Marcuse.

the results reported by Marcuse[2] (using two different approximations); our results (for $0 < kd_1 < 3$) are shown by the solid curve. As expected, the radiation losses generally decrease with increasing waveguide thickness. We observe that all three results show very good agreement for kd_1 values in the range between 0.7 and 1.7. For values of kd_1 below 0.7, our results show somewhat higher radiation losses. For kd_1 values above 1.7, Marcuse's two results begin to diverge; one of the approximations yields meaningless negative radiation losses for kd_1 values greater than 2.0, while the other approximation shows higher radiation losses than those found by our approach. The radiation loss minimum appears to occur at $kd_1 \approx 2.7$. Marcuse shows a loss minimum of about 2%; our results show a loss minimum below 0.5%.

Joint Services Technical Advisory Committee
F44620-74-C-0056

E-W. Hu and L. Bergstein

REFERENCES

1. L. Bergstein, E.W. Hu and S.T. Peng, "Modal Analysis of Surface Waveguide Discontinuity," Progress Report No. 40 to JSTAC, Polytechnic Inst. of New York, Report No. R-452.40-75, 66-74 (1975).

2. D. Marcuse, "Radiation Losses of Tapered Dielectric Slab Waveguide," Bell System Tech. J., 273-290 (1970).

STRAY FIELD OF A SLOTTED MICROSTRIP LINE

I. Palócz

In a number of recent computer memories, microstrip-type lines are used. In an attempt to gain a better understanding of the behavior of these lines, the present author conducted several previous studies.[1,2,3] In this contribution, the magnetic field at the ground plane and at the air boundary is investigated. This calculation is performed in connection with the stray fields occurring in memories. It should be mentioned that, in present day experimental memories, the so-called bit lines are very close to the ground plane, and the stray magnetic fields do represent a limit in packaging the elements closer and closer: the effective area of one storage element is determined by this field.

In this study, we calculate the magnetic field in the x direction, which is of practical interest. For notation, see Figure 1. Note that H_x is proportional to E_y and that the proportionality factor is $(\varepsilon_o/\mu_o)^{1/2}$. The calculation is a straightforward extension of this author's previous calculation,[2] where the current distribution has been analyzed by using a finite Hilbert transformation.[4] Since the current distribution is known, the magnetic field is simply given by (see Eq. (36) of Reference 2):

$$H_x = \frac{h}{\pi} \int_a^1 i(x') \left[\frac{1}{(x-x')^2 + h^2} + \frac{1}{(x+x')^2 + h^2} \right] \tag{1}$$

where the notation is evident from Figure 1. (By multiplying all the dimensions by w/2 one obtains results for the line of width w, rather than for the line of width "two.")

The slotted line will now be compared with an unslotted one (see the right upper part of Figure 1.) For the latter, the ground plate is infinite and the upper electrode is semi-infinite, and the coordinate system is so chosen that x equals zero at the edge of the semi-infinite electrode. The stray field of this line has long ago been calculated in a rather ingenious manner by Hermann Von Helmholtz and Gustav Kirchhoff.[5]

The parametric representation of the coordinate x and y are given by

$$\frac{x}{b} = \frac{1}{\pi}(1 - t - e^{-t} \cos s)$$

$$\frac{y}{b} = \frac{1}{\pi}(s - e^{-t} \sin s) \tag{2}$$

Letting s = const., we obtain the equipotential lines, while t = const. yields the lines of force.

The potential of the upper plate is zero (s = 0); it is evident from Eq. (2) that for

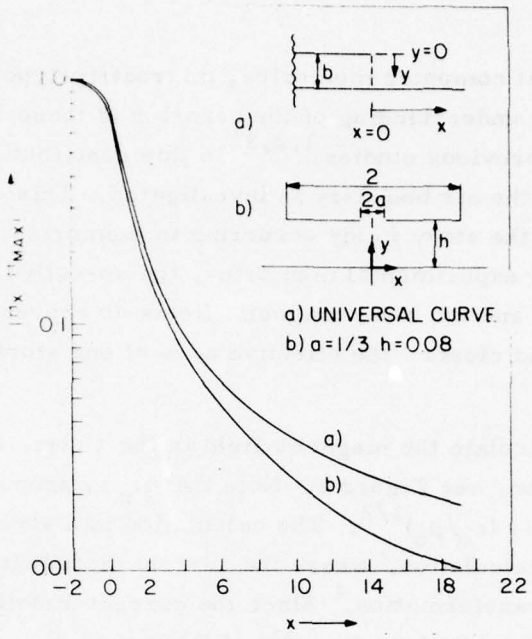


Fig. 1. a) H_x versus x for a semi-infinite strip above ground; b) H_x versus x/b for a slotted line.

this case y = 0. For t = 0, x = 0, and for all positive and negative values of t the value of x is negative; the electrode is semi-infinite. For the ground plane $s = \pi$; as evident from Eq. (2) x assumes all the negative and positive values as t varies from $-\infty$ to ∞. Since the field strength at the ground plane is of interest, one writes for the neighborhood of the ground plate $s = \pi - \Delta s$, where Δ is small; omitting quadratic and higher order terms in the expansion of cos Δs and sin Δs, one gets

$$\frac{x}{b} = \frac{1}{\pi}(1 - t - e^{-t}) \tag{3a}$$

$$\frac{y}{b} = 1 - \frac{\Delta s}{\pi}(1 + e^{-t}) \tag{3b}$$

Hence the magnitude of the field strength H_x, normalized to H_{max}, is

$$\frac{H_x}{H_{max}} = \frac{1}{H_{max}} \frac{\Delta s}{\Delta y} = \frac{1}{1 + e^{-t}} \tag{4}$$

For any value of t, one gets x/b from Eq. (3a) and the corresponding normalized field from Equation (4). The result of this calculation is shown in the curve a) of Figure 1. The slotted case is plotted in curve b) of Fig. 1 for the case of a = 1/3, h = 0.08.

Joint Services Technical Advisory Committee
F44620-74-C-0056

I. Palócz

REFERENCES

1. H. Guckel, D.A. Brennan and I. Palócz, "A Parallel-Plate Waveguide Approach to Microminiaturized, Planar Transmission Lines for Integrated Circuits," IEEE Trans. MTT, MTT-15, No. 8, 468-475 (August 1967).

2. I. Palócz, "The Integral Equation Approach to Currents and Field in Plane-Parallel Transmission Lines," J. of Math. and Mechanics, 15, No. 4, 541-560 (April 1966).

3. M.E. Hellman and I. Palócz, "The Effects of Neighboring Conductors on the Currents and Fields in Plane Parallel Transmission Lines," IEEE Trans. MTT, MTT-17, No. 5, 254-259 (May 1969).

4. F.G. Tricomi, "On the Finite Hilbert Transformation," Quart. J. Math., Oxford, Ser. 2, 2, 199-211 (1951).

5. a) G. Kirchhoff, Monats. d. Akad. d. Wissenschaften, Berlin, 144 (March 1877).
   b) H.V. Helmholtz, Berl. Monetsber, April 23, 1868; Phil. Mag. (November 1868).

PHYSICAL PROCESS OF SAW SCHOTTKY DIODE MEMORY CORRELATOR

W-C. Wang

It is attempted here to explain the physical process involved in SAW diode memory correlators, based on the published experimental result of Ingebrigtsen, Cohen and Mountain.[1] They obtained their correlation read-out by observing the variations of convolution output at frequency 2ω between two r.f. pulses, one long and one short, at frequency ω. Figure 2 of Ref. 1 is redrawn here as Figure 1. It shows the Schottky

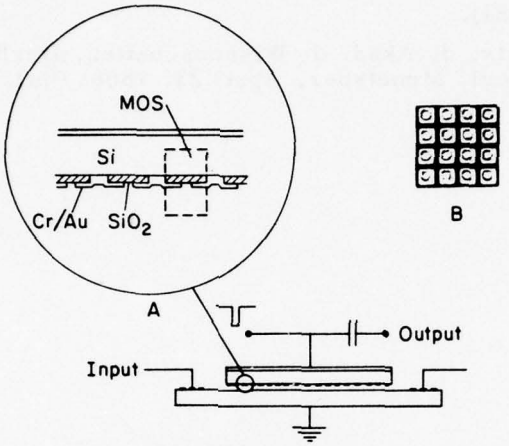


Fig. 1. Schematic of the experimental configuration with a photograph of the Schottky-diode matrix.

diode matrix (inserts A and B) and the diode - $LiNbO_3$ memory structure.

The spherical platinum silicide Schottky diodes are of 5 μm in diameter. The periodicity between the diodes is 12.7 μs. Each diode is overlayed with a square of Cr/Au with dimensions $\approx 10 \times 10\ \mu m^2$.

For clarity, we also single out the MOS portion of the device which is indicated by the dotted square and redrawn in insert A.

The MOS portion of the structure is found to be essential in the operation of the diode memory correlator. This point will become clearer as we proceed.

The amplitude variation of the convolved signal versus applied d.c. plate pulses of short duration (150 nsec) obtained by Ingebrigtsen et al., is redrawn in Fig. 2, where the plate pulse is used to forward bias the Schottky diodes. When the Schottky diode is forward biased, the injected carriers will distribute themselves over the entire Cr/Au layer ($10 \times 10\ \mu m^2$) and serve as a reverse bias for the MOS structure (insert A).

The shape of the convolution curve in Fig. 2 can be qualitatively explained based on the theory developed by Gautier and Kino for reverse-biased MOS structure.[2,3]

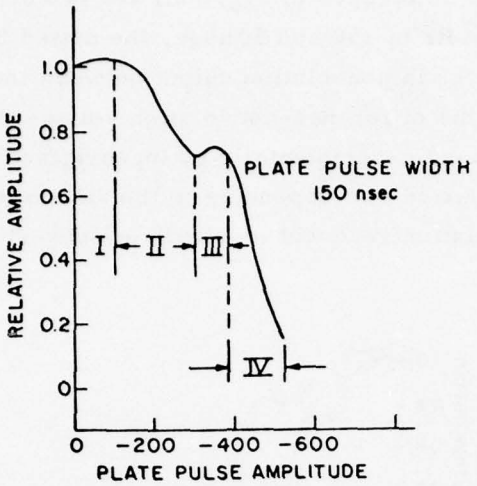


Fig. 2. Normalized convolution output as a function of plate pulse voltage. (Figure 3 of Reference 1.)

Let us assume that when no bias voltage is applied, the n-type MOS structure is at the flat band condition. When a small voltage is applied to the structure, as shown by portion I of the curve, the depletion width widens and the convolution output increases. The reason is that the sonic induced field can now better modulate the depletion width. When the depletion width further increases as shown in portion II of the curve, it effectively increases the gap spacing between the semiconductor and the piezoelectric substrate, thus it decreases the piezoelectric coupling and decreases the convolution output as well. The portion III of the curve is the part stated by Ingebrigtsen et al. to be most sensitive to the voltage variations. It is the place where the population of holes starts to surpass the population of electrons, i.e., the inversion layer begins to form there. When the inversion takes place, the transverse d.c. acoustoelectric voltage reverses its sign and begins to forward bias the MOS diode. Its net effect slightly reduces the reverse-biasing voltage and results in a slight increase in the convolution output. When the reverse-bias still increases further as it is shown in portion IV of the curve, it thickens the inversion layer. Since the inversion layer will shield the semiconductor from the piezoelectric field, the convolution output decreases quite rapidly.

The explanation on the convolution output given so far is made under the assumption that the MOS diode is under continuous reverse biasing, but no initial charges are stored on the Cr/Au film. Supposing that there are initial charges present there, then in effect the stored electrons would increase the reverse-bias voltage by ΔV_S. That is, the stored charges would shift the set of convolution curves obtained by Ingebrigtsen

et al. to the left by ΔV_S. The solid lines in Fig. 3(a) are two of their original convolution curves with plate pulse width of 150 and 50 nsec, the dotted lines represent the shifted curves. The differences in convolution output between the two sets of curves are plotted in Figure 3(b). This difference-convolution voltage Δv_c should be equal to the correlation read-out obtained experimentally by Ingebrigtsen et al., since the ΔV_S shift in Fig. 3(a) has been adjusted corresponding to the amount of charge stored in their experiment. The correlation read-out (normalized) observed by Ingebrigtsen et al. is redrawn in Figure 3(c).

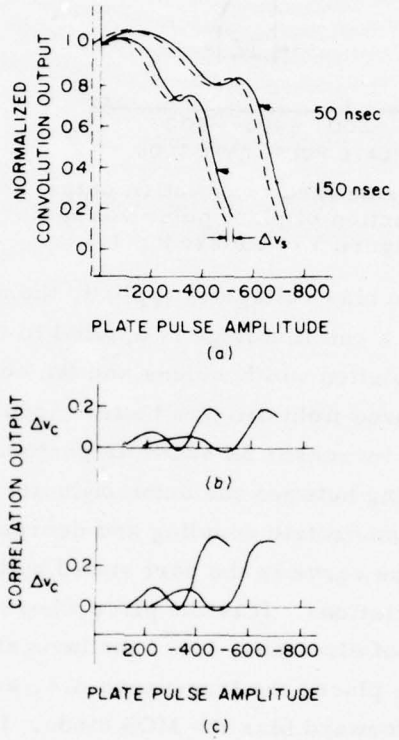


Fig. 3. (a) and (c) are the plots of convolution output and correlation dip as a function of plate pulse height. The stored pulse and the reading pulse length are of 250 nsec long. (Figure 3 of Reference 1). Figure 3(b) plots the differences in convolution outputs (shown in 3(a)) between the solid curves and the dotted curves.

In comparing Figs. 3(b) and 3(c), one observes the apparent agreement. The differences in magnitude between Figs. 3(b) and 3(c) are introduced by the fact that they have normalized their experimental data. Excellent agreement has been obtained by taking into account the normalization factor.

In conclusion, we have shown that the MOS portion of the structure does play an essential part in the diode memory correlator. This conclusion is expected to be general and applies to other published types of diode memory systems as well. The

Schottky diode serves as a carrier injector and the amount of charge injected is modulated by the acoustic write-in signal. As far as the correlation read-out is concerned, the contribution of the Schottky diode to the convolution signal is small, possibly due to the simple fact that it only occupies one-eighth of the total surface area.[4] It is also clear that in this particular read-out system the maximum correlation output appears at the inversion region, which would not be the case for other types of read-out systems.[5-8] In the second paper of Ingebrigtsen and Stern that have modified their memory structure, replacing the Cr/Au overlay by a high-resistive polysilicon film. The modified structure is more complicated to analyze, since the contribution of the resistive polysilicon layer may not be ignored.

Joint Services Technical Advisory Committee
F44620-74-C-0056

W-C. Wang

REFERENCES

1. K.A. Ingebrigtsen, R.A. Cohen and R.W. Mountain, Appl. Phys. Lett., 26, 596 (1975).
2. H.R. Gautier, M.L. Report No. 2448, Stanford University (June 1975).
3. S. Takada, K. Hoh, H. Hayakawa and N. Mikoshiba, Supplement to the J. of the Japan Soc. of Appl. Phys., 42, 21 (1973).
4. The contribution of the convolved signal from Schottky-diodes in general should decrease as the reverse-biasing voltage is increased due to the effective increase of gap spacing between the semiconductor and the piezoelectric substrate.
5. K.A. Ingebrigtsen and E. Stern, Appl. Phys. Lett., 27, 170 (1975).
6. C. Maerfeld, Ph. Defranould and P. Tournois, Appl. Phys. Lett., 27, 577 (1975).
7. W.C. Wang, 1975 Ultrasonic Symposium Proceedings, Los Angeles, California.
8. A.G. Bert and G. Kantorowicz, Electronics Lett., 11, 523 (1975).

EXPERIMENTS ON SAW DIODE CORRELATOR

W-C. Wang

Utilizing nonlinear interaction of acoustic surface waves to perform the function of correlation has been under intensive study for several years. Advances made in this period have been very rapid, especially in the direction of the memory correlator.[1-3] We have done some experimental studies on the Si-$LiNbO_3$ diode correlator. Most of the experimental results presented here were obtained, using Thomson-CSF vidicon diode arrays, while the author was on leave at Thomson-CSF. Similar results have been obtained at the Polytechnic using RCA 4532 vidicon diode mosaic. We describe first the inter-relationships among convolutuon, stored charge and correlation, and then the effect of d.c. pulse biasing on the correlation output.

A. Inter-relationships Among Convolution, Correlation and Charge Storage

Figure 1 is the experimental configuration which is the same as that reported by

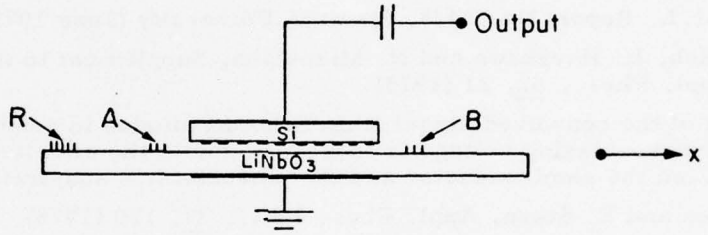


Fig. 1. Our experimental configuration. Center frequencies of A and B, 60 MHz and R, 120 MHz.

Maerfeld et al. P-n diode arrays are used in the structure. A brief description on the diode mosaic is given here. The original Si wafer is of 10 Ω-cm resistivity and n-type. On its surface a 5000 A SiO_2 layer is grown. Then through an etching process, arrays of 5 μm diameter holes are made on the SiO_2 film. At the hole sites the p-n diode arrays are formed. The carrier density on the P-region is about 10^{16} atoms/cm^3. The diodes are then overlayed with an $8 \times 8\,(\mu m)^2$, P^+ Si layer of 1 μm thick. The periodicity of the diodes is 12.5 μm.

On the $LiNbO_3$ substrate, as shown in Fig. 1, a pair of 60 MHz and one 120 MHz interdigital transducers are deposited at their respective locations A, B and R.

Let a signal of frequency ω be applied to the transducer A exciting a surface wave propagating in the x direction, charge would be stored due to the action of piezoelectric field associated with the surface wave forward-biasing the diode arrays. The amount of charge stored can be controlled by many factors such as the diode characteristics,

the pattern of the diode arrays and their overlays, the signal amplitude, the d-c biasing condition and the repetition rate of the applied signals. If another signal of the same carrier frequency ω is applied to transducer B exciting a wave propagating in the -x direction, then due to nonlinear mixing, two signals are produced. One is the commonly known convolution signal at the sum frequency 2ω and the other at the difference frequency, which is independent of time but with a spatial variation of $\exp(2kx)$. Both the linear and nonlinear field induced by the two oppositely propagating waves will affect the amount and the pattern of the charge storage. However, the concern here is the retention of both the amplitude and phase information of the spatially varying nonlinear signal, since it is the signal which will later interact with reading-pulse to produce correlation output. Thus, it is important to know the conditions under which one can optimize the storage of a spatially varying signal.

The stored charge will act as a reversed bias for both the diode arrays and the MOS portion of the structure. It has been known that the convolution signal is sensitive to the biasing condition, i.e., by observing the variation of convolution output, one may obtain the information on charge storage. This point has been confirmed by the following experiment.

Write-in signals of large amplitudes 20 volts, p-p are first applied to transducers A and B. The observed convolution output is quite large, actually out of scale (Fig. 2) and then, at time t_1 the signal amplitudes are suddenly switched from 20 volts to 6 volts, the convolution output is observed to drop to level I at t_1 and then slowly increase to level II. This rising time constant can be related to the leakage rate of the stored charge. Under the same experimental condition, but 1.2 msec after writing signals applied to A and B, a low level reading-pulse is applied to transducer R(120 MHz) to observe the changes in correlation output corresponding to the sudden change of writing signal amplitudes applied at transducers A and B. Figure 2(b) shows the changes as a function of time. At t_1 the correlation output starts to decrease, and the time it requires to reach steady-state is seen to be closely related to that in Figure 2(a). In fact, they have the same time constants. Figures 3(a) and 3(b) are obtained under the same condition as Figs. 2(a) and 2(b) except that the ambient temperature is $9^{\circ}C$ for Fig. 3 and $18^{\circ}C$ for Figure 2. It is noted that the time constant at low temperature is larger than that at high temperature, which is expected from the temperature characteristic of a reversed biased diode.

It is of importance to write down the correlation output we have observed in the above experiment. Let the signals applied to transducers A and B be $A e^{j\omega t}$ and $B e^{j\omega t}$, then during the write-in process the stored signal is

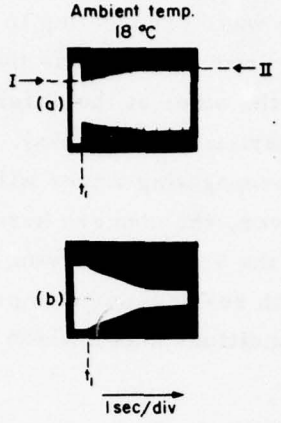


Fig. 2. (a) Convolution and (b) Correlation, as a function of time.

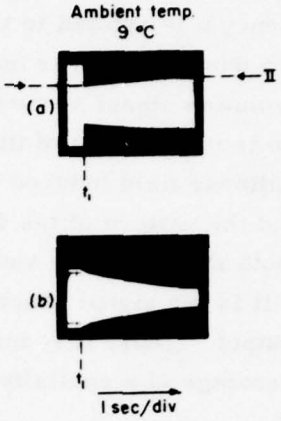


Fig. 3. (a) Convolution and (b) Correlation, as a function of time.

$$P_1 \int_0^T A(t - \frac{x}{v})\, B(t + \frac{x}{v} - \frac{L}{v})\, e^{j(2kx - kL)} dt$$

$$= P_1 e^{j(2kx - kL)} \int A(\tau + \frac{L - 2x}{v})\, B(\tau) d\tau = C(\frac{L - 2x}{v})\, e^{j(2kx - kL)} \qquad (1)$$

After the application of the read-out signal, $Re^{j2\omega t}$, one obtains the following correlation output,

$$P_2 e^{jkL} \int C(\frac{L - 2x}{v})\, R(t - \frac{x}{v})\, e^{j2\omega t}\, d(\frac{x}{v}) \qquad (2)$$

where P_1 and P_2 are constants, L is the path length between transducers A and B, and T represents the duration of the shorter signal.

Equation (1) indicates that signal A has correlated with signal B through time integration during the process of the charge storage.

Equation (2) indicates that the reading signal R has correlated with the stored signal, $C(2x/v)$ through spatial integration during the read-out process. The output represented by Eq. (2) is a complicated one. However, it can be simplified and has practical importance under the following two conditions.

(1) If signal R is of very short duration and can be approximated by $\delta(t)$, then the read-out signal truly represents the correlation between signals A and B. However, it should be pointed out that since the time integration is performed by charge accumulation, the time rate of charge build-up is required to be linear. Otherwise, discrepancies will be introduced.

(2) If the signal B is of very short duration, $C(-2x/v) \simeq A(-2x/v)$, then the read-out signal represents the correlation between signals A and R. Here the correlation process is performed through spatial integration. It should also be pointed out, however, that a time scaling factor of one-half is introduced in this type of reading and writing process. The time scaling factor can be removed by using layered structure, $LiNbO_3$-Si-BGO.

For the experimental result presented in this write-up, neither the duration of signal R nor the duration of signal B is made very short, i.e., the correlation output described here is a complicated one. The reason for not using a very short reading or writing pulse is due to the fact that we prefer to have large signal output in conducting various studies of this device.

The write-in process adopted here is believed to be the most efficient one, since the writing pulse is applied to the transducer and its amplitude does not have to be very large in comparison with other write-in methods used by Bers and Cafarella, by Ingebrigtsen, Stern, Mountain and Cohen.

B. Effect of d-c Pulse Biasing During Writing and Reading Intervals

We have observed that both the signal amplitudes and the rate of charge build-up are affected, when a d-c pulse is applied. The extent of the effect has been found to be strongly dependent on the writing and reading pulse amplitudes, and also on the time and duration during which the d-c bias is applied. The first case we will consider is when the d-c pulse is applied during the time that the charge is being stored, and then we will consider the effect of the d-c bias when it is applied during the time that the correlation is being read out.

The triangular shaped signal indicated by I in Fig. 4 is the convolution output between two rectangular pulses, both of 4 μsec in duration and 4 volts in amplitude. The pulse of large amplitude corresponds to the reading pulse radiated through the air. The correlation output which should appear after the reading pulse at position II is too small to be recognized. When a pulsed d-c voltage is applied to forward-bias the diode array during the time when the charge storage is in process, shown in Fig. 4(b), the correlation output is greatly enhanced. In fact, its amplitude is comparable to that of the convolution signal. The observed enhancement is due to the fact that in Fig. 4(a) the writing signal is too small to give appreciable charge storage, but in Fig. 4(b) with the help of forward biasing the operating point on the diode characteristic is shofted and more charge is stored. It is also noted that the convolution output is increased in Figure 4(b).

Figure 5 describes another facet of the experiment. It indicates the effect of forward biasing on the time rate of charge build-up. Here the repetition rate of the writing signal is less than 3 pps, which is much slower than the reading pulse rate $\simeq$ 100 pps.

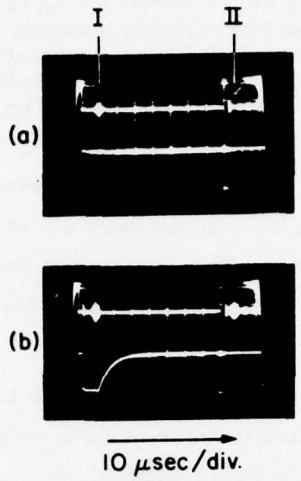


Fig. 4. Effect of d-c pulse biasing on the signal outputs. Position I is convolution and II is correlation. Input signal amplitudes 4^{V}. Reading pulse height 8^{V}. (a) No d-c bias applied. (b) Forward-bias applied during writing process. Scale 100^{V}/sec.

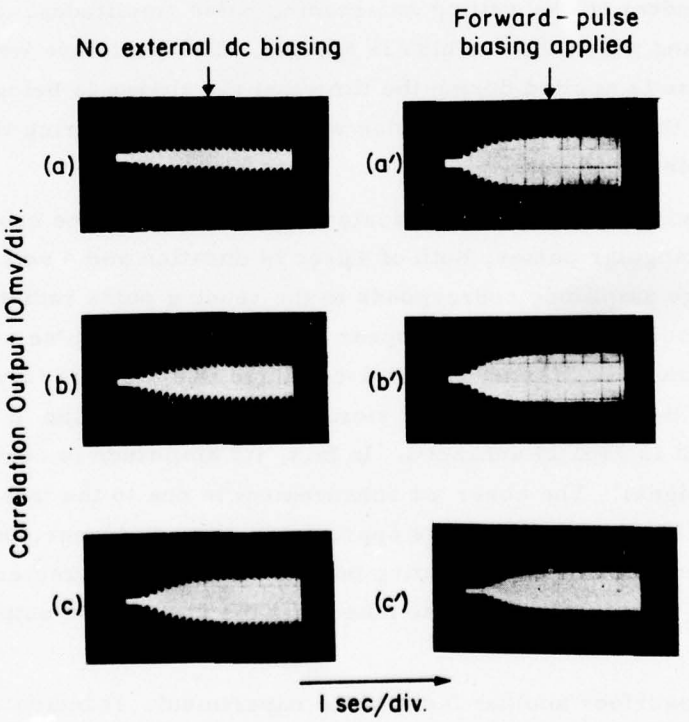


Fig. 5. Effect of forward biasing during writing process on the time rate of charge build-up.

Oscillograms a, b and c in Fig. 5 correspond to the condition at which no d-c bias is applied. Oscillograms a′, b′ and c′ are taken under the condition that the diodes are forward biased when the writing process is in progress. The effect of reading pulse height on the rate of charge build-up is also observed. The reading pulse heights in oscillograms a and a′ are of 18 volts, in b and b′, 14 volts and in c and c′, 11 volts. We notice that the common feature in the three sets of oscillograms is the disappearance of the fast time decaying constant when forward-biasing is applied. This is because more charge is stored when the diode is forward biased. When the reading pulse is applied, due to the presence of the stored charge the diode operating point is shifted to the lower part of the diode characteristic and gives a much longer time constant. This special feature is expected to be useful when multiple writings are employed or when the correlation process is performed through time integration.

Now let us consider the case where the d-c bias is applied at the time interval when the correlation output is observed, as indicated in Figure 6. The extent of the

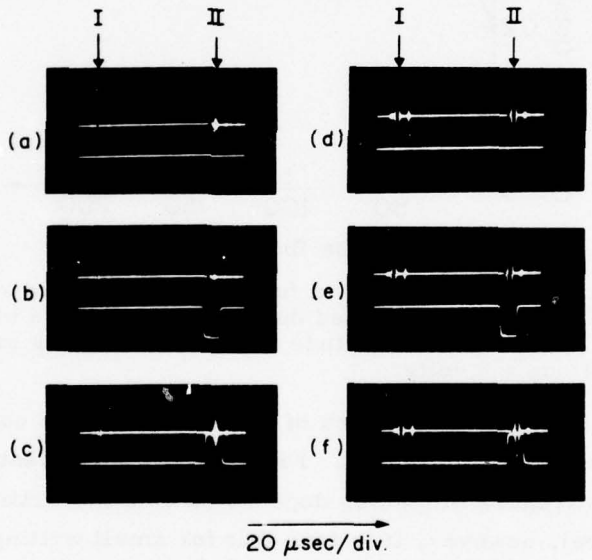


Fig. 6. Effect of d-c pulse biasing (applied during read-out interval) on the signal outputs, Position I indicates convolution output and II correlation output.

biasing effect depends on the amplitude of writing signal. The amplitude of writing signal for oscillograms a, b and c is of 9 volts. The first triangular signal corresponds to the convolution output; the last signal corresponds to the correlation read-out. We observe that the correlation signal is reduced in Fig. 6(b) when the structure is forward-

biased and enhanced in Fig. 6(c) when the structure is reversely biased. Oscillograms d, e and f are obtained under the same condition as that of a, b and c except that the writing signal amplitude is increased to 22 volts. In this case we noted that the correlation output is only slightly affected by the d-c biasing. We further noted that the effect of forward-biasing increases the correlation output and the effect of reverse-biasing decreases the correlation output, which is contrary to what we have observed in oscillograms a, b and c.

In Fig. 7 we plot the amplitude of correlation read-out as a function of the reverse-

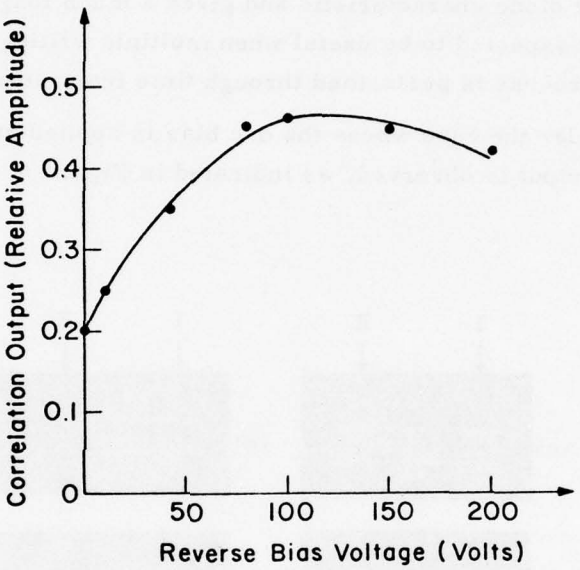


Fig. 7. Correlation output as a function of reverse d-c biasing. Bias voltage is applied during read-out time interval. Writing signal amplitude = 6 volts. Reading signal amplitude = 9 volts.

bias voltage for writing signal amplitudes of 6 volts. Here the correlation output first reaches a maximum and then decreases. From these experiments of Figs. 6 and 7 it is clear that the effectiveness of biasing depends on both the writing and reading signal amplitudes. In general, however, it is true that for small writing signal and moderate reading strength, the correlation output vs. reverse bias height follows the curve described in Figure 7.

The author wishes to thank C. Maerfeld for discussions and for using his experimental set-ups.

Joint Services Technical Advisory Committee
F44620-74-C-0056

W-C. Wang

REFERENCES

1. K.A. Ingebrigtsen, R.A. Cohen and R.W. Mountain, Appl. Phys. Lett., 26, 596 (1975).
2. K.A. Ingebrigtsen and E. Stern, Appl. Phys. Lett., 27, 170 (1975).
3. C. Maerfeld, Ph. Defranould and P. Tournois, Appl. Phys. Lett., 27, 577 (1975).

RAYLEIGH WAVE FIELDS EXCITED BY AN ACOUSTIC BEAM INCIDENT FROM A LIQUID NEAR THE RAYLEIGH CRITICAL ANGLE

Y. L. Hou and H. L. Bertoni

An acoustic beam reflected into a liquid from the surface of a solid at or near the angle of phase matching with the free-surface Rayleigh wave exhibits features not explained by geometrical acoustics.[1,2,3] The beam center is shifted from the geometrical optics position and its profile has variations different from that of the incident beam. Based on the properties of the reflected beam, it has been established that these effects result from the excitation and re-radiation of the Rayleigh wave, which is of the leaky type due to the presence of the liquid.[4]

Although it is known indirectly that the incident beam in the fluid excites the Rayleigh wave, the elastic fields in the solid have, in fact, never been computed. The work described here establishes that the elastic fields are in the form of a Rayleigh wave whose amplitude varies along the surface. Besides completing the description of the beam shift phenomena, this study allows computation of the power flow in the solid when viscous absorption or loss is present. Previously, power flow was computed only for lossless solids using conservation of energy arguments. In the lossless case it was found that up to 80 percent of the incident power could be transferred to the Rayleigh wave prior to re-radiation.

The power flow in the solid, and the affect of loss on it, are of interest in nondestructive evaluation (NDE) of solids. Loss in polycrystalline materials is strongly influenced by the grain size.[5] Thus, loss measurements can be used to determine grain size. Grain size is of particular interest in structural materials (aluminum, steel, etc.) since it is a major factor determining the materials' strength. Detecting grain size near the surface of a solid from observation of the phase of reflected plane waves has been proposed as an NDE technique.[6,7] Simpler amplitude measurements for bounded beam reflection have also been shown to be applicable to NDE determination of grain size.[8,9] Knowledge of the power flow, in the solid due to a bounded incident beam, helps to understand the features of the reflected beam that would be used in NDE amplitude measurements.

A. Approximations for Plane Wave Transmission and Reflection Coefficients

In the presence of a liquid at the surface of a solid, the free surface Rayleigh wave is perturbed into a leaky wave, for the usual case when the wavenumber k in the liquid is greater than the shear and longitudinal wavenumbers k_s and k_p in the solid. For $\exp(-i\omega t)$ time dependence, the leaky wave fields vary along the surface shown in Fig. 1 as $\exp[i k_\ell x]$ where

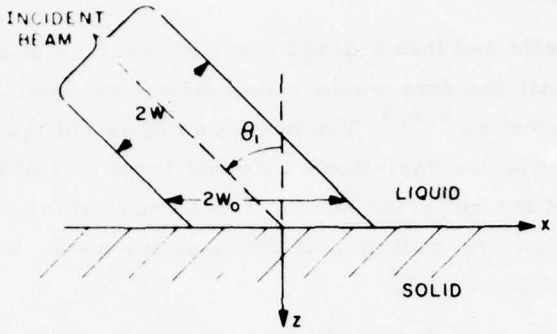


Fig. 1. Gaussian beam incident on a liquid-solid interface.

$$k_\ell = \beta + i\alpha \tag{1}$$

Here, β is the phase constant of the leaky wave and is close to the phase constant k_R of the free-surface Rayleigh wave. The attenuation constant α of the leaky wave results from the fact that the Rayleigh wave sheds power into the liquid at an angle $\theta_\ell = \sin^{-1}(\beta/k)$, as it propagates along the surface.

Because the leaky Rayleigh wave is a resonant solution satisfying boundary conditions at the interface, the wavenumber k_ℓ will appear as a pole in the plane wave transmission coefficients T_p and T_s, and in the reflection coefficient R. Let $k_x = k \sin\theta$ be the transverse wave number of a plane wave incident at the angle θ. The transmission coefficients, as a function of k_x, may be expanded about the pole at $k_x = k_\ell$ in a Laurent series. Retaining only the first term then gives a good approximation to $T_p(k_x)$ and $T_s(k_x)$ for k_x in the vicinity of β.

The approximation described above takes the form

$$T_p(k_x) \approx \frac{N_p(k_\ell)}{D'(k_\ell)} \frac{1}{k_x - k_\ell}$$

$$T_s(k_x) \approx \frac{N_s(k_\ell)}{D'(k_\ell)} \frac{1}{k_x - k_\ell} \tag{2}$$

where

$$N_p(k_x) = 2k_s(\kappa_s^2 - k_x^2)$$

$$N_s(k_x) = 4k_s^2 \kappa_p k_x \tag{3}$$

are the numerators of T_p and T_s with

$$\kappa_s(k_x) = \sqrt{k_s^2 - k_x^2}$$

$$\kappa_p(k_x) = \sqrt{k_p^2 - k_x^2} \tag{4}$$

Also $D'(k_x)$ is the derivative of the denominator

$$D(k_x) = 4k_x^2\kappa_s\kappa_p + (\kappa_s^2 - k_x^2)^2 + \frac{\rho_f}{\rho} k_s^4 \kappa_p/\kappa_f \tag{5}$$

with

$$\kappa_f(k_x) = \sqrt{k^2 - k_x^2} \tag{6}$$

and ρ_f, ρ being the mass densities of the liquid and solid, respectively. Since k_ℓ is close to k_R, the Rayleigh wave number, we may make the approximation of replacing k_ℓ in $N_p(k_\ell)$, $N_s(k_\ell)$ and $D'(k_\ell)$ by k_R. Thus Eq. (2) becomes

$$T_p(k_x) \approx \frac{N_p(k_R)}{D'(k_R)} \frac{1}{k_x - k_\ell}$$

$$T_s(k_x) \approx \frac{N_s(k_R)}{D'(k_R)} \frac{1}{k_x - k_\ell} \tag{7}$$

A simple Laurent expansion cannot be used for the reflection coefficient R. For a lossless solid $|R| = 1$ for $k_x > k_s$. Thus R must have a zero k_0 located at the point $k_0 = k_\ell^*$. The correct form found for R is[4]

$$R = \frac{k_x - k_0}{k_x - k_\ell} \tag{8}$$

When loss is present in the solid $k_0 \neq k_\ell^*$. Loss may be introduced into the computation of k_ℓ and k_0 by allowing the longitudinal and shear wavenumbers to have complex values. It has been found that the loss tangent of the longitudinal wave is not significant in determining k_0 and k_ℓ. For simplicity we therefore take both loss tangents to be the same, so that

$$k_{p,s} = k'_{p,s}(1 + i\,d/2\pi) \tag{9}$$

where d is referred to as the loss factor.

The variation in the pole and zero location in the complex k_x plane as a function of $d = d_p = d_s$ are shown in Fig. 2 for a water-stainless steel interface. For $d = 0$ we have that $k_0 = k_\ell^*$, while for $d \neq 0$ both k_0 and k_ℓ move upward as d increases -- the numbered points on the curves refer to particular values of d. It is seen that at some critical value d_c of d, the zero crosses the real axis. Thus, a real-axis zero is produced which corresponds to total suppression of the geometrically reflected field. Since the loss factor d depends on frequency, the value $d = d_c$ will be obtained at a particular frequency. The existence of a frequency at which the reflection is a minimum has previously been observed, and is referred to as the frequency of least reflection.[6,7] Thus the approximation of Eq. (8) for the reflection coefficient explains in a simple way the dependence of R on loss for plane waves incident near the Rayleigh angle $\theta_r \approx \theta_\ell$. The plot of Fig. 2 has the same general features as those previously reported for an aluminum-water interface,[8] although the shape of the curves is somewhat different owing to the differences in mass density.

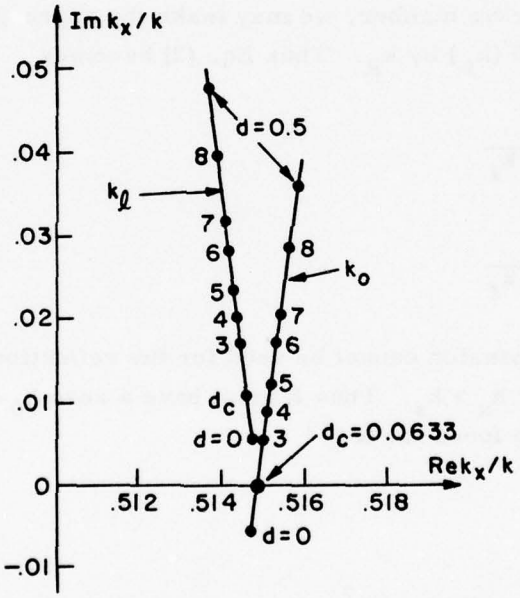


Fig. 2. Loci of the pole k_ℓ and zero k_0 of $R(k_x)$ in the complex k_x plane for a water-stainless steel interface with loss factor d as parameter.

As part of the study reported here, we have investigated the accuracy of the approximation in Eq. (8) by comparing numerical results from Eq. (8) with the exact expression for the reflection coefficient. The deviation of $\Delta\phi$ in the phase of the approximate expression in Eq. (8) from the exact expression for R is plotted in Fig. 3 as a

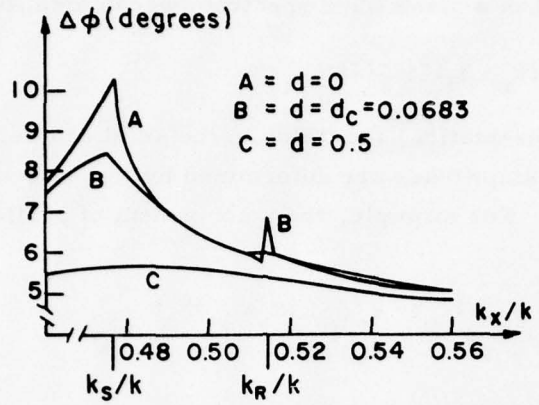


Fig. 3. Phase deviation $\Delta\phi$ (equal to the exact minus the approximate phase of R) as a function of k_x in the vicinity of the Rayleigh wavenumber.

function of k_x normalized to the fluid wavenumber k for a water-stainless steel interface. Curves for the cases of no loss (d = 0), critical loss ($d = d_c = 0.0683$) and high loss (d = 0.5) all show that $\Delta\phi$ is small for transverse wavenumbers k_x somewhat greater than the shear wavenumber k_s of the solid. Moreover, $\Delta\phi$ varies slowly with k_x, except when k_x is close to k_p in the case of critical loss, so that the relative error is even smaller. The value of $|R|$, as determined from Eq. (8), is also close to the exact value, so that Eq. (8) represents a simple but accurate approximation for R.

B. Fields Excited in the Solid by a Bounded Beam

For the realistic case of a bounded beam incident on the surface, the incident field is composed of a spectrum of plane waves distributed over a range of angles about the angle of incidence θ_i of the beam axis. As a result of the variation of R, T_p and T_s with $k_x = k \sin\theta$, for θ_i near θ_ℓ, the profile of the reflected beam and the variation of the substrate fields along x will differ substantially from that of the incident beam. In addition, the presence of loss can be expected to influence the beam profile and x variation, as well as the amplitude.

To study the effect of loss in the solid, we consider an incident beam having a Gaussian profile. The x component of particle velocity is assumed at the interface z = 0 to have the form

$$v_i(x, 0) = \exp(-x^2/w_0^2)\exp(ik_i x) \qquad (10)$$

where $k_i = k \sin\theta_i$ and w_0 is the half-width of the beam as measured parallel to the interface (see Figure 1).

The Gaussian beam has a plane wave spectrum whose amplitude is given by

$$V(k_x) = \sqrt{\pi}\, w_0 \exp[-(k_x - k_i)^2 w_0^2/4] \tag{11}$$

Using the plane wave representation, the fields in the solid can be found as a superposition of plane waves whose amplitudes are determined by Eq. (10) and the plane wave transmission coefficients. For example, the x component of particle velocity in the solid is given by

$$v_x(x,z) = \frac{\rho_f}{\rho} \int_{-\infty}^{\infty} V(k_x)\left[T_p e^{i\kappa_p z} - \frac{\kappa_s}{k_x} T_s e^{i\kappa_s z}\right] e^{ik_x x}\, dk_x \tag{12}$$

Similar expressions hold for the other field components.

For wide beams, the spectrum of plane waves in Eq. (12) is concentrated about $k_x = k_i$. For k_i close to β, and hence to k_R, use of the approximations in Eq. (7) for T_p and T_s is justified. The plane wave fields in the solid for k_x near k_R, i.e., for $k_x > k_s$, will have κ_p and κ_s imaginary, as seen from Equation (4). Thus $v_x(x,z)$ will decay exponentially into the solid, so that we focus attention on observation points near the surface. For these observation points the terms $\exp(i\kappa_p z)$ and $\exp(i\kappa_s z)$ will be slowly varying functions of k_x over the significant part of the plane wave spectrum, and may therefore be treated as constants in the integration. Similarly, the factor κ_s/k_x may be treated as a constant.

Evaluating the terms taken as constants in the integration at $k_x = k_R$ we obtain for Eq. (12)

$$v_x(x,z) = \frac{\rho_f}{\rho}\left[\frac{N_p(k_R)}{D'(k_R)} e^{-|\kappa_p(k_R)|z} - \frac{\kappa_s(k_R)}{k_R}\frac{N_s(k_R)}{D'(k_R)} e^{-|\kappa_s(k_R)|z}\right] E(x) \tag{13}$$

where

$$E(x) = \sqrt{\pi}\, w_0 \int_{-\infty}^{\infty} \frac{e^{-(k_x - k_i)^2 w_0^2/2}}{k_x - k_\ell} e^{ik_x x}\, dk_x \tag{14}$$

The integral in Eq. (14) may be evaluated in closed form, and is found to be

$$E(x) = \frac{i\sqrt{\pi}}{2} w_0 e^{\gamma^2} \operatorname{erfc}(\gamma)\, v_i(x,0) \tag{15}$$

where $\operatorname{erfc}(\gamma)$ is the complementary error function and

$$\gamma = -\frac{x}{w_0} + i(k_i - k_\ell) w_0/2 \tag{16}$$

The polarization and z dependence of the velocity in Eq. (13) is that of a Rayleigh wave. The variation of the fields with x results from the excitation and re-radiation into the fluid of the Rayleigh wave. Similarly, the other field components can be shown to be those of a Rayleigh wave whose amplitude varies as E(x).

The maximum magnitude achieved by $|E(x)|$ occurs when $k_i = \beta$ so that γ in Eq. (16) is real. For $|k_i - \beta| w_0 > \pi/2$, the maximum value of $|E(x)|$ decreases rapidly with increasing phase deviation $|k_i - \beta|$. Thus, strong field excitation occurs in the substrate only when the incident beam is close to phase matching with the Rayleigh wave.

C. Power Flow in the Solid

Using Eq. (14), we have computed the total power $P_x(x)$ carried along x, per meter along y, in the Rayleigh wave fields by integrating the local Poynting vector over z from z = 0 to z = ∞. Thus

$$P_x(x) = \int_0^\infty -\frac{1}{2} \mathrm{Re}[\underline{v}(x,z) \cdot \underset{\sim}{T}^*(x,z) \cdot \underline{x}_0]\, dz \qquad (17)$$

where $\underset{\sim}{T}(x,z)$ is the stress tensor and * represents complex conjugate. It is found that $P_x(x)$ is proportional to $|E(x)|^2$.

The power $P_x(x)$ is plotted in Fig. 4 as a function of x/w_0 for a water-stainless steel interface with a beam half-wdith $w_0 = 0.68/\alpha_0$ and several values of the loss parameter d. Since the Rayleigh wave excitation is strongest when $k_i = \beta$, the computation made in arriving at Fig. 4 assumes the angle of incidence is such as to satisfy this relationship for d = 0. In Fig. 4, $P_x(x)$ has been normalized to the total power in the incident beam. The quantity α_0 is the attenuation constant of the leaky Rayleigh wave for the lossless case d = 0, and the value of w_0 is chosen such that $P_x(x)$ takes on its highest possible value of 0.8.

To the left of the maximum of $P_x(x)$, which occurs at some value $x = x_m$, the power in the Rayleigh wave increases as the wave is excited. Subsequently its amplitude decreases due to radiation into the fluid, and due to acoustic damping for $d \neq 0$. The maximum power in the Rayleigh wave is seen to decrease as loss increases. This effect is more clearly shown in Fig. 5, where the value of $P_x(x_m)$, normalized to the power in the incident beam, is plotted as a function of d for several values of normalized beam width $\alpha_0 w_0$. The power $P_x(x_m)$ decreases monotonically with d, and for d = 0 has a maximum value when the width of the incident beam is such that $\alpha_0 w_0 = 0.68$.

From Fig. 4 it is seen that for each curve, the maximum slope $|dP_x/dx|$ to the left of the peak decreases as the loss factor d increases. However, for $d < d_c$, the maximum of $|dP_x/dx|$ to the right of the peak is almost independent of d. These

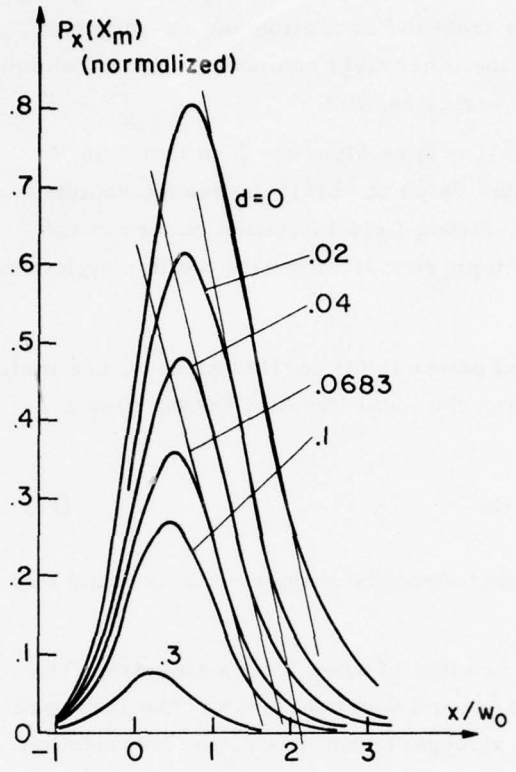


Fig. 4. Variation of the power $P_x(x)$ carried by the Rayleigh wave with distance along the interface and loss factor d as a parameter.

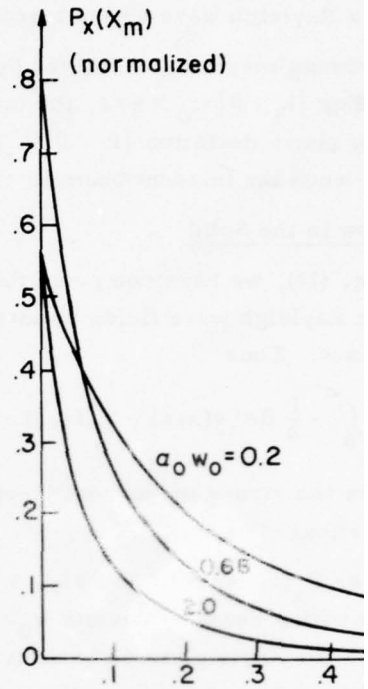


Fig. 5. Variation of the peak $P_x(x_m)$ carried by the wave with loss factor normalized half-width a parameter.

observations suggest that for $d < d_c$, the presence of loss interferes with th of the leaky wave, but does not significantly increase the attenuation consta leaky wave. This effect is confirmed by the small movement of the pole k_ℓ for $d < d_c$. For $d > d_c$, the maximum value of $|dP_x/dx|$ decreases with d of the peak, indicating that loss has a significant effect on α, as well as on tion of the leaky wave. The foregoing discussion is consistent with the inte of the dependence of the reflected beam profile on loss.[8]

National Science Foundation
ENG 75-00569

Y. L. Hou and H.

REFERENCES

1. W.G. Neubauer, "Ultrasonic Reflection of a Bounded Beam at Rayleigh Angles for a Plane Liquid-Solid Interface," J. Appl. Phys., 44, 48-55

2. M.A. Breazeale and L. Adler, "Reflection of a Gaussian Ultrasonic Beam From a Liquid-Solid Interface," J. Acoust. Soc. Am., 56, 866-872 (1974).

3. T.J. Palona, M. Behravesh and W.G. Mayer, "Ultrasonic Bounded Beam Reflection Effects at a Liquid-Solid Interface," J. Acoust. Soc. Am., 56, 1773-1775 (1974).

4. H.L. Bertoni and T. Tamir, "Unified Theory of Rayleigh-Angle Phenomena for Acoustic Beams at Liquid-Solid Interfaces," Appl. Phys., 2, 157-172 (1973).

5. E.P. Papadakis, "Revised Grain-Scattering Formulas and Tables," and "Ultrasonic Attenuation Caused by Scattering in Polycrystalline Metals," J. Acoust. Soc. Am., 37, 703-717 (1965).

6. F.L. Becker and R.L. Richardson, "Influence of Material Properties on Rayleigh Critical-Angle Reflectivity," J. Acoust. Soc. Am., 51, 1609-1617 (1972).

7. F.L. Becker, "Phase Measurement of Reflected Ultrasonic Waves Near the Rayleigh Critical Angle," J. Appl. Phys., 42, 199-202 (1971).

8. H.L. Bertoni, Y.L. Hou and T. Tamir, "Application of Rayleigh Angle Phenomena to the Detection of Acoustic Loss in Solids," *Progress Report* No. 39 to JSTAC, Polytech. Inst. of New York, Report No. R-452.39-74, 40-49 (1974).

9. Y.L. Hou and H.L. Bertoni, "Evaluation of Substrate Properties Using an Acoustic Beam in a Prism Coupled Via a Fluid Layer," *Progress Report* No. 40 to JSTAC, Polytech. Inst. of New York, Report No. R-452.40-75, 82-93 (1975).

RAYLEIGH WAVE SCATTERING FROM A LIQUID-FILLED CRACK: PART I, VARIATIONAL EXPRESSIONS FOR REFLECTION AND TRANSMISSION COEFFICIENTS

Y. L. Hou and H. L. Bertoni

In the work reported here, a variational expression was developed giving the reflection and transmission coefficients of a Rayleigh wave at a liquid-filled crack. The expression gives the amplitude and phase of these coefficients, as well as the power scattered into the bulk waves. Rayleigh wave scattering from cracks is a fundamental problem in elasticity, and is of interest for microwave acoustics and nondestructive evaluation (NDE), which are areas of active research at the Polytechnic, as well as seismology.

Unwanted cracks on the surface of a crystal will scatter the Rayleigh wave in a SAW device and degrade its performance. A knowledge of the dependence of scatter on crack size would permit formulation of criteria for maximum defect size. Alternatively, machined cracks, or narrow grooves, could be used as reflecting elements in RAC-type devices. Here it is essential to know both the reflection coefficient and the power scattered into bulk waves in order to design and evaluate devices.

An important problem in NDE is the location of cracks initiating at the surface of a part. Since the initial size of the crack determines the working life of a part, it is important to find cracks above a certain size. One type of device proposed for locating cracks employs an elastic beam in a prism coupled to the solid under test by a liquid layer or an acoustic beam incident from a liquid bath.[1] When incident at the angle for excitation of the Rayleigh wave, the reflected beam contains features that are strongly perturbed when the Rayleigh wave is scattered by a crack. Monitoring the features of the reflected beam can then be used to locate the crack.

In this initial study the crack is assumed to have slip boundary condition, wherein the normal components of stress and particle motion are continuous, but the tangential component of stress vanishes. These boundary conditions were chosen because of interest in the NDE problem described above in which a liquid is present at the crack. In the case of a thin, unstressed fatigue crack, the slip boundary conditions may be more applicable than the free-surface boundary conditions. In any case, the slip boundary condition leads to the simplest variational expression.

The crack is assumed to be oriented perpendicular to the direction of propagation of the Rayleigh wave. Both cracks initiating at the surface and located wholly within the solid are considered. While the depth into the solid is finite, the length of the crack parallel to the surface is taken infinite, so that the scattering problem is two-diemsnional. While considerable literature exists on the scattering of bulk waves by

cracks (see survey in Ref. 2), Rayleigh wave scattering by a crack at or near the surface of a half-space does not seem to have been treated previously.

Starting with an integral representation for the fields inside a solid in terms of the fields at its surface, an integral equation is obtained for the jump in the tangential component of particle motion across the crack. The variational expression is then obtained as an approximate solution to the integral equation. The trial function in the variational expression is the jump in the tangential component of motion.

In Part II of this two-part report, choice of the trial functions is discussed and numerical results are obtained for the reflection and transmission coefficients. To obtain convergence of the integrals, it is found necessary to include the correct edge condition at the crack tip. Trial functions are therefore taken as the product of the field dependence of the Rayleigh wave, and a function accounting for the edge condition.

A. Integral Equation for Scattering by a Crack

The geometry of the slip crack is shown in Fig. 1(a) for the case of a crack originating at the surface, and in Fig. 1(b) for crack inside the solid. A Rayleigh wave is incident from $x = -\infty$, and both it and the crack have infinite extent along y. For generality we consider the crack inside the solid, and recover the crack at the surface by setting $d = 0$.

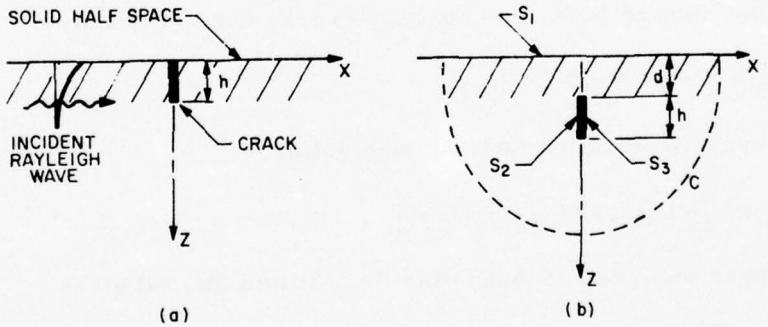


Fig. 1. Crack geometry for: (a) crack originating at surface; and (b) crack entirely within solid, indicating surfaces of integration.

Let $G_{Ki}(x, z; x', z')$ represent the two-dimensional half space Green's function, i.e., the $i = 1, 3$ component of particle displacement due to a point source oriented along the $K = 1, 3$ axis and having harmonic time dependence $\exp(j\omega t)$. The corresponding stress field is denoted by $T_{ij}^{(K)}(x, z; x', z')$. Further, let $U_i(x, z)$ be the $i = 1, 3$ component

of particle displacement of the field scattered by the crack, i.e., the total field less the incident Rayleigh wave field. The stress field of the scattered wave is $T_{ij}(x, z)$.

The Green's function fields and the scattered fields satisfy the dynamic equations. Thus, for any closed cylindrical surface parallel to the y axis and having a generating curve S(x, z), it can be shown that[3]

$$U_K(x, z) = \oint [G_{Ki}(x', z'; x, z)\, T_{ij}(x', z') - U_i(x', z')\, T_{ij}^{(K)}(x', z'; x, z)]\, n_j\, ds(x', z') \tag{1}$$

Here integration is taken over the closed curve $S(x', z')$ and n_j represents the $j = 1, 3$ component of the outward unit normal. Summation over repeated indices is implied.

The curve S is taken as the free surface S_1 in Fig. 1(b), the semi-cylinder C at infinity and the sides S_2 and S_3 of the crack. On S_1, the stress T_{ij} and $T_{ij}^{(K)}$ vanish, while on C all field quantities vanish, if small loss is assumed to insure convergence. At the slip crack the 13 component of the total stress must vanish. Therefore $T_{13}(x, z)$ is the negative of the 13 stress component of the incident Rayleigh wave, which is continuous across the crack. Thus $T_{13}(x, z)$ is continuous across the crack, as is $T_{11}(x, z)$ and $U_1(x, z)$. Similarly G_{Ki} and $T_{ij}^{(K)}$ are continuous across the crack, since they apply to a semi-infinite soild without a crack. Only $U_3(x, z)$ is discontinuous.

Using the forging properties of the fields at the surfaces S_1, S_2, S_3 and C, the representation theorem in Eq. (1) reduces to

$$U_K(x, y) = \int_d^{d+h} T_{31}^{(K)}(0, z'; x, z)\, J_3(z')\, dz' \tag{2}$$

where $J_3(z')$ is the jump in $U_3(x, z)$ across the crack, and is given by

$$J_3(z') = U_3(0^+, z') - U_3(0^-, z') \tag{3}$$

From the dynamic equation for isotropic solids,

$$T_{13}(x, z) = \mu\left[\frac{\partial}{\partial x} U_3(x, z) + \frac{\partial}{\partial z} U_1(x, z)\right] , \tag{4}$$

where μ is the shear modulus. Substituting Eq. (2) into Eq. (4) gives

$$T_{13}(x, z) = \int_d^{d+h} A(z'; x, z)\, J_3(z')\, dz' \tag{5}$$

where

$$A(z'; x, z) = \mu \frac{\partial}{\partial z} T_{31}^{(1)}(0, z'; x, z) + \mu \frac{\partial}{\partial x} T_{31}^{(3)}(0, z'; x, z) \tag{6}$$

As discussed above, at the crack $x = 0$, $d < z < d+h$, the stress component $T_{13}(0, z)$ is the negative of $T^I_{13}(0, z)$ for the incident Rayleigh wave, which is given by

$$T^I_{13}(0, z) = -2\mu k_R \kappa_{pR} \left[e^{-j\kappa_{pR} z} - e^{-j\kappa_{sR} z} \right] \tag{7}$$

Here k_R is the Rayleigh wave number and κ_{pR} and κ_{sR} are the values of the wavenumbers

$$\kappa_p(k_x) = \sqrt{k_p^2 - k_x^2}$$
$$\kappa_s(k_x) = \sqrt{k_s^2 - k_x^2} \tag{8}$$

evaluated at $k_x = k_R$, with k_p and k_s the longitudinal and shear wavenumbers of the solid.

Evaluating Eq. (5) at $x = 0$, $d < z < d+h$ we therefore find

$$-T^I_{13}(0, z) = \int_d^{d+h} A(z'; 0, z)\, J_3(z')\, dz' \tag{9}$$

In Eq. (9), T^I_{13} and $A(z'; 0, z)$ are known, while the jump $J_3(z')$ in the particle displacement $U_3(z')$ is unknown. Equation (9) thus represents an integral equation for the unknown jump. This equation is used below to obtain a variational expression for the Rayleigh wave reflection and transmission coefficients.

B. Fourier Representation for A and the Relation Between R and T

From a consideration of the spectral representation of $A(z', x, z)$ it is possible to find a relationship between the Rayleigh wave reflection and transmission coefficients R and T. The spectral representation is then required to find a variational expression that gives R and T.

The stress components $T^{(1)}_{31}$ and $T^{(3)}_{31}$ in Eq. (6) can be found from Eq. (4) with U_i replaced by appropriate components G_{Ki} of the Green's function. Thus $A(z'; x, z)$ consists of combinations of derivatives with respect to x, x', z and z' of the Green's function $G_{Ki}(x', z'; x, z)$. The Green's function, and hence $A(z'; x, z)$, can be separated into particular and homogeneous parts $A_p(z'; x, z)$ and $A_h(z'; x, z)$. The particular portion corresponds to radiation by the point source into an infinite medium. The homogeneous portion corresponds to the fields generated by reflection of the point source radiation at the free surface $z = 0$.

The homogeneous portion of the Green's function[4] leads to

$$A_h(z';x,z) = \frac{j\mu^2}{4\pi\rho\omega^2}\int_{-\infty}^{\infty}\Big[c(k_x)\,e^{-j\kappa_p(z+z')} + e(k_x)\,e^{-j\kappa_s(z+z')} + d(k_x)\,e^{-j(\kappa_p z+\kappa_s z')} + d(k_x)\,e^{-j(\kappa_s z+\kappa_p z')}\Big]\frac{e^{-jk_x x}}{F(k_x)}\,dk_x \qquad (10)$$

where

$$\begin{aligned} c(k_x) &= -4k_x^2\kappa_p\left[(2k_x^2 - k_s^2)^2 - 4k_x^2\kappa_s\kappa_p\right] \\ d(k_x) &= 8k_x^2\kappa_p(2k_x^2 - k_s^2)^2 \\ e(k_x) &= \frac{1}{\kappa_s}(2k_x^2 - k_s^2)^2\left[(2k_x^2 - k_s^2)^2 - 4k_x^2\kappa_s\kappa_p\right] \\ F(k_x) &= 4k_x^2\kappa_s\kappa_p + (2k_x^2 - k_s^2)^2 \end{aligned} \qquad (11)$$

The integration in Eq. (10) is taken along the real k_x axis, as shown in Figure 2. The location of the branch points, branch cuts and Rayleigh poles are indicated for the case of small loss in the solid.

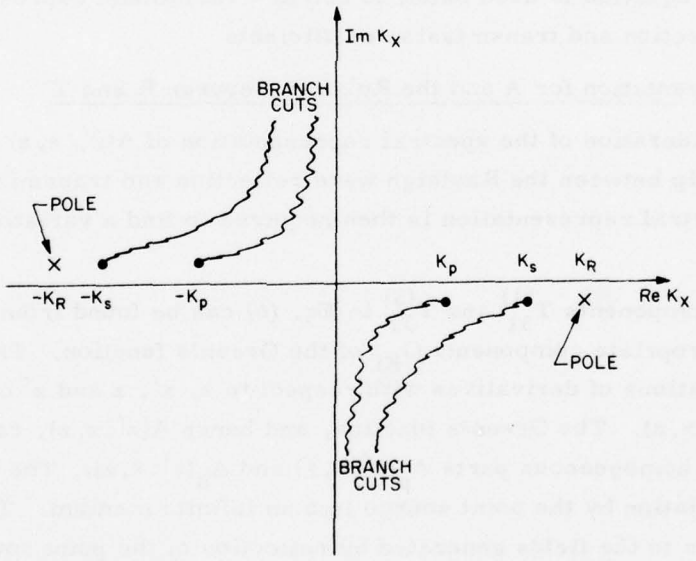


Fig. 2. Integration path along the Re k_x axis in the complex k_x plane, indicating branch points, branch cuts and poles of $A_h(z';x,z)$.

For $x < 0$, which is to the left of the crack, the integration path in Fig. 2 can be deformed into the upper half plane leaving a residue contribution and the contribution

A_h^{bc} coming from integration around the branch cuts originating at $-k_s$ and $-k_p$. Substituting the particular branch cut and residue portions of A into Eq. (5) it can be shown that for $x < 0$,

$$T_{13}(x,z) = \int_d^{d+h} \left[A_p(z';x,z) + A_h^{bc}(z';x,z) \right] J_3(z')\, dz' + \frac{\mu}{4\rho\omega^2} \frac{e^{jk_R x}}{F'(k_R)} c(k_R)(I_p - I_s) \left[- \frac{T_{13}^I(0,z)}{k_R \kappa_{pR}} \right] \tag{12}$$

where $F'(k_R)$ is the derivative of $F(k_x)$ evaluated at $k_x = k_R$ and

$$I_p = \int_d^{d+h} e^{-j\kappa_{pR} z'} J_3(z')\, dz' \tag{13}$$

and I_s is similarly defined with κ_{sR} replacing κ_{pR}. In deriving Eq. (12), the equalities $e(k_R) = c(k_R) = -d(k_R)$ were used.

The left-hand side of Eq. (12) represents the scattered field and contains two portions; one portion being the bulk wave and the other the reflected Rayleigh wave $-R\,T_{13}^I(0,z)\,e^{jk_R x}$. Equating this reflected Rayleigh wave field with the portion on the right-hand side of Eq. (12) having the Rayleigh dependence gives

$$-R = -\frac{\mu}{4\rho\omega^2} \frac{c(k_R)(I_p - I_s)}{F'(k_R)\, k_R \kappa_{pR}} \tag{14}$$

For $x > 0$, that is to the right of the crack, the integration path in Fig. 2 can be deformed into the lower half plane. As a result, A_h will be the sum of a residue contribution from the pole at k_R and the integration around the branch cuts originating at k_p and k_s. When substituted into Eq. (5), one obtains the same expression as given in Eq. (12) with k_R replaced by $-k_R$. The left-hand side of the new equation will consist of the bulk wave contribution to the scattered field plus the Rayleigh wave term $(T-1)\,T_{13}^I(0,z)\,e^{-jk_R x}$. Here T is the transmission coefficient, and when the foregoing Rayleigh wave term is added to the incident Rayleigh wave one obtains the transmitted Rayleigh wave. From arguments similar to that used to obtain Eq. (14), one finds

$$T - 1 = -\frac{\mu}{4\rho\omega^2} \frac{c(k_R)(I_p - I_s)}{F'(k_R)\, k_R \kappa_{pR}} \tag{15}$$

where the symmetries $c(-k_R) = c(k_R)$ and $F'(-k_R) = -F'(k_R)$ have been used. Comparing Eqs. (14) and (15) one finds the relation

$$R = 1 - T \tag{16}$$

between R and T.

C. Variational Expression for R and T

To obtain the desired variational expression, we specialize Eq. (12) to the case $x = 0$, which is equivalent to using the Fourier representation for A in Equation (9). Since $T_{13}(0,z)$ is equal to $-T^I_{13}(0,z)$, and using Eq. (14), we obtain from Eq. (12)

$$-T^I_{13}(0,z) = \int_d^{d+h} \left[A_p(z';0,z) + A_h^{bc}(z';0,z)\right] J_3(z')\,dz' - R\,T^I_{13}(0,z) \tag{17}$$

We now multiply Eq. (17) by $J_3(z)$ and integrate from d to d+h. With the help of Eq. (7), the foregoing integration yields

$$(R-1)(-2k_R\kappa_{pR})(I_p - I_s) = \iint_d^{d+h} \left[A_p(z';0,z) + A_h^{bc}(z';0,z)\right] J_3(z')\,J_3(z)\,dz'\,dz \tag{18}$$

To complete the derivation, we divide Eq. (18) by $(I_p - I_s)^2$ and use Eq. (14) to eliminate $1/(I_p - I_s)$ from the left-hand side of the resulting equation. By this process we obtain

$$\frac{1-R}{R}\,\frac{\mu^2 c(k_R)}{2\rho\omega^2 F'(k_R)} = \frac{1}{(I_p - I_s)^2} \iint_d^{d+h} \left[A_p(z';0,z) - A_h^{bc}(z';0,z)\right] J_3(z')\,J_3(z)\,dz'\,dz \tag{19}$$

Equation (19) is the variational expression from which we obtain R. It is variational in the sense that the variation of the right-hand side with respect to $J_3(z)$ vanishes when the right-hand side is evaluated at the exact solution for $J_3(z)$. Proof of this statement is straightforward, but is not given here.

The importance of Eq. (19) is that given reasonable guess for $J_3(z)$, we obtain a more accurate value for R. Note that $c(k_R)$ and $F'(k_R)$ are known functions so that R can be determined once the right-hand side of Eq. (19) has been found numerically. Since $J_3(z)$ appears twice in the numerator and in the demoninator, multiplication of $J_3(z)$ by a constant will not change the value of R. Evaluation of the integrals in Eq. (19) and the appropriate choice of $J_3(z')$ is discussed in Part II of this two-part report.

National Science Foundation
ENG75-00569

Y. L. Hou and H. L. Bertoni

REFERENCES

1. Y. L. Hou and H. L. Bertoni, "Evaluation of Substrate Properties Using an Acoustic Beam in a Prism Coupled via a Fluid Layer," Progress Report No. 40 to JSTAC, Polytech. Inst. of New York, Report No. R-452.40-75, 82-92 (1975).
2. E. A. Kraut, "Review of Theories of Scattering of Elastic Waves by Cracks," IEEE Trans., SU-23, 162-167 (1976).
3. L. Knopoff, "Diffraction of Elastic Waves," JASA, 28, 217-229 (1956).
4. R. W. Fredericks, "Scattering of Elastic Pulses by Obstacles of Infinite Impedance and Semi-Infinite Dimensions on the Surface of a Half-Space," Ph. D. Thesis, U. of Calif., Los Angeles (1959).

RAYLEIGH WAVE SCATTERING FROM A LIQUID-FILLED CRACK: PART II, EDGE CONDITIONS AND NUMERICAL RESULTS

Y. L. Hou and H. L. Bertoni

In Part I of this two-part report, we have discussed the motivation for studying the reflection of Rayleigh waves from cracks, which is a fundamental problem of elasticity with application to SAW devices, NDE and seismology. The study carried out here applies to scattering from infinitely long cracks that are perpendicular to the direction of propagation and have slip boundary conditions. The slip boundary condition applies to a liquid filled crack, which is of particular interest in NDE. The analysis in Part I leads to a variational expression from which the reflection coefficient R and transmission coefficient T could be found. Further, the fraction of the incident power scatter into bulk waves is found as $1 - |R|^2 - |T|^2$.

The variational expression contains a double integral of a product consisting of a kernel, known in terms of its Fourier representation, and a trial function, which is the jump $J_3(z)$ across the crack of the particle displacement component parallel to the crack. In this report, it is shown that the double integral can be reduced to a single integral over wavenumbers, provided a suitable choice of $J_3(z)$ is made. The choice of $J_3(z)$ must take into account the field behavior near the edge of the crack required by the physical principle of finite energy storage.

With the correct choice of $J_3(z)$, the resulting wavenumber integral can be shown to converge. Numerical data obtained from numerical integration of the wavenumber integral is presented. The data is for a crack originating at the surface and covers the range from shallow to deep cracks. It shows only weak dependence on Poisson's ratio for the solid and a surprisingly large amount of power scattered into bulk waves, even for deep cracks.

A. Simplification of Integrals for Numerical Evaluation

In Part I, the variational expression for R was obtained in the form

$$\frac{1-R}{R}\,\frac{\mu^2 c(k_R)}{2\rho\omega^2 F'(k_R)} = \frac{N}{(I_p - I_s)^2} \tag{1}$$

where

$$N = \int_d^{d+h}\int \left[A_p(z';0,z) - A_h^{bc}(z';0,z)\right] J_3(z') J_3(z)\, dz'\, dz \tag{2}$$

and the other terms in Eq. (1) are defined in Eqs. (11) and (13) of Part I. The integrals I_p and I_s are single integrals over finite limits, and can easily be evaluated. However,

A_h^{bc} in Eq. (2) is given by the wavenumber integral in Eq. (10) in Part I taken over the contour P in the complex k_x plane shown in Figure 1.

The function $A_p(z';0,z)$ comes from the Green's function for an infinite medium, and can therefore be represented in terms of various derivatives of the Hankel functions. Alternatively, it may be expressed as

$$A_p(z';0,z) = \int_{-\infty}^{\infty} a(\kappa)\, e^{-j\kappa(z-z')} d\kappa \tag{3}$$

where

$$a(\kappa) = \frac{-1}{k_{xs}}\left[4k_{xp}k_{xs}\kappa^2 + (k_s^2 - 2\kappa^2)^2\right] \tag{4}$$

and

$$k_{xp} = \sqrt{k_p^2 - \kappa^2}$$
$$k_{xs} = \sqrt{k_s^2 - \kappa^2} \tag{5}$$

Using Eq. (3) and the appropriate expression for A_h^{bc}, Eq. (2) can be written as

$$N = \iint_d^{d+h} \int_{-\infty}^{\infty} a(\kappa)\, e^{-j\kappa(z-z')} J_3(z) J_3(z')\, d\kappa\, dz\, dz'$$
$$+ \iint_d^{d+h} \int_P \Big\{ c(k_x) e^{-j\kappa_p(z+z')} + e(k_x) e^{-j\kappa_s(z+z')}$$
$$+ d(k_x)\left[e^{-j(\kappa_p z + \kappa_s z')} + e^{-j(\kappa_p z' + \kappa_s z)}\right]\Big\} \frac{1}{F(k_x)} J_3(z) J_3(z')\, dk_x\, dz\, dz' \tag{6}$$

Interchanging orders of integration in Eq. (6) gives

$$N = \int_{-\infty}^{\infty} a(\kappa) I(\kappa) I(-\kappa)\, d\kappa + \int_P \left[c(k_x) I^2(\kappa_p) + 2\, d(k_x) I(\kappa_p) I(\kappa_s) + e(k_x) I^2(\kappa_s)\right] \frac{1}{F(k_x)}\, dk_x \tag{7}$$

where

$$I(\kappa) = \int_d^{d+h} J_3(z)\, e^{-j\kappa z}\, dz \tag{8}$$

Note that in Eq. (1), $I_p = I(\kappa_p)$ and $I_s = I(\kappa_s)$. Thus, if the form of $J_3(z)$ is sufficiently

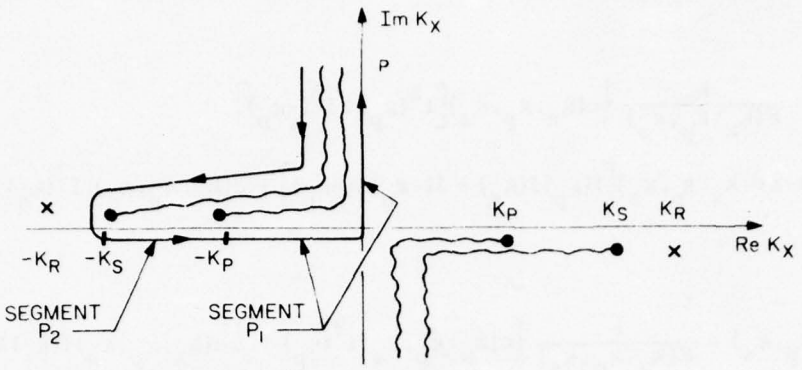


Fig. 1. Branch points, branch cuts and the deformed integration path in the complex k_x plane.

simple so that the integral in Eq. (8) can be carried out analytically, then only a single numerical integration in Eq. (7) over wavenumbers must be carried out in order to determine the right hand side of Equation (1).

The path P may be divided into four parts. Let P_1 be the part running from $-k_p$ to $k_x = 0$ and then along the Im k_x axis to infinity. In the limit of zero loss, the branch cuts in Fig. 1, defined by Im $\kappa_p = 0$ and Im $\kappa_s = 0$, coincide with portions of the Re k_x axis and with the Im $k_x > 0$ axis. In this limit, the portion P_4 of the path lying on the opposite side of the branch cuts to P_1 will also coincide with portions of the Re k_x axis and the Im $k_x > 0$ axis. We take P_2 as the segment $-k_s < k_x < -k_p$ and P_3 as the portion of P opposite to P_2 across the cut.

The functions $c(k_x)$, $d(k_x)$, $e(k_x)$ and $F(k_x)$ in Eq. (7) are explicitly functions of κ_p and κ_s, as seen from Eq. (11) of Part I. These functions have the symmetry property

$$\begin{aligned} c(k_x; -\kappa_p, -\kappa_s) &= -c(k_x; \kappa_p, \kappa_s) \\ d(k_x; -\kappa_p, -\kappa_s) &= -d(k_x; \kappa_p, \kappa_s) \\ e(k_x; -\kappa_p, -\kappa_s) &= -e(k_x; \kappa_p, \kappa_s) \\ F(k_x; -\kappa_p, -\kappa_s) &= F(k_x; \kappa_p, \kappa_s) \end{aligned} \tag{9}$$

where the dependence on κ_p and κ_s has been displayed explicitly. Using these symmetry properties and parts of P described above, Eq. (7) can be written as

$$N = \int_{-\infty}^{\infty} a(\kappa) I(\kappa) I(-\kappa)\, d\kappa + \int_{P_1} n_1(k_x)\, dk_x + \int_{-k_s}^{-k_p} \left[n_2(k_x; \kappa_p, \kappa_s) - n_2(k_x; \kappa_p, -\kappa_s) \right] dk_x \tag{10}$$

where

$$n_1(k_x) = \frac{1}{F(k_x;\kappa_p,\kappa_s)}\left\{c(k_x;\kappa_p,\kappa_s)\left[I^2(\kappa_p)+I^2(-\kappa_p)\right]\right.$$
$$\left. + 2\,d(k_x;\kappa_p,\kappa_s)\left[I(\kappa_p)I(\kappa_s) + I(-\kappa_p)I(-\kappa_p)\right] + e(k_x;\kappa_p,\kappa_s)\left[I^2(\kappa_s)+I^2(-\kappa_s)\right]\right\} \tag{11}$$

and

$$n_2(k_x;\kappa_p,\kappa_s) = \frac{1}{F(k_x;\kappa_p,\kappa_s)}\left\{c(k_x;\kappa_p,\kappa_s)I^2(\kappa_p) + 2\,d(k_x;\kappa_p,\kappa_s)I(\kappa_s)I(\kappa_p)\right.$$
$$\left. + e(k_x;\kappa_p,\kappa_s)I^2(\kappa_s)\right\} \tag{12}$$

Along the path P_1, the wavenumber κ_p varies from 0 to ∞ and is real. Letting

$$\kappa_p = \kappa$$
$$k_x = \sqrt{k_p^2 - \kappa^2} \tag{13}$$

we may convert the integration over P_1 into an integral over $0 \le \kappa < \infty$ by change of variable. Noting from Eq. (4) that $a(\kappa)$ is an even function, Eq. (10) becomes

$$N = \int_0^{\infty}\left\{2\,a(\kappa)I(\kappa)I(-\kappa) + n_1\left[k_x(\kappa)\right]\frac{dk_x}{d\kappa}\right\}d\kappa + \int_{-k_s}^{-k_p}\left[n_2(k_x;\kappa_p,\kappa_s) - n_2(k_x;\kappa_p,-\kappa_s)\right]dk_x \tag{14}$$

The second integral in Eq. (14) is over finite limits and the integrand is well behaved so that the integration can be carried out numerically. The convergence of the integral over infinite limits in Eq. (14) is discussed in the next section.

B. Edge Condition and the Choice of $J_3(z)$

For the remainder of this report we specialize our discussion to the case of a crack originating at the surface, so that d = 0, and penetrating into the solid to a depth h. At the surface z = 0, the two sides of the crack are free to shear or slide past each other so that $J_3(0)$ is finite. However, at the tip of the crack (x = 0, z = h) the two sides of the crack are constrained by the surrounding medium so that no shear is possible and hence $J_3(h) = 0$.

In order for the energy stored in the vicinity of the edge to be finite, the particle displacement must vary as the one half (or higher power) of distance from the edge. Thus $J_3(z)$ must vary as $(h - z)^q$, where $q \ge 1/2$, near the edge.[1,2] Aside from the required variation near the edge, $J_3(z)$ should have variation with z that is similar to the

v_3 component of particle displacement of the Rayleigh wave. Thus we assume the form

$$J_3(z) = f(z)\left[e^{-j\kappa_{sR}z} - \frac{2k_R^2 - k_s^2}{2k_R^2} e^{-j\kappa_{pR}z}\right] \tag{15}$$

where f(z) is the function accounting for the edge condition.

We have considered two different choices for f(z). The first choice was

$$f(z) = \sqrt{h - z} \tag{16}$$

Using Eq. (16) in Eq. (15), the integration in Eq. (8) can be obtained in closed form in terms of complex error functions. For $\kappa \to \infty$, it is found that $I(\kappa)$ varies as $1/\kappa$. Using this it can be shown that the integrand in the first integral of Eq. (14) varies as $1/\kappa^3$ for $\kappa \to \infty$. Thus the infinite integral converges. If $J_3(z)$ did not vanish at $z = h$, then the infinite integral would not converge, so that the edge condition is essential in obtaining a variational solution.

The second choice of f(z) used was

$$f(z) = \cos\left(\frac{\pi}{2h} z\right) \tag{17}$$

which varies linearly near the edge $z = h$. This choice of f(z) gives a simpler expression for $I(\kappa)$ than does Equation (16). The infinite integral in Eq. (14) is also found to converge for f(z) in Eq. (17), and the computation time is much less than needed when Eq. (16) is used. The numerical results obtained for the reflection and transmission coefficients R and T using Eqs. (16) and (17) are close to each, indicating that the variational expression is not sensitive to the choice of $J_3(z)$

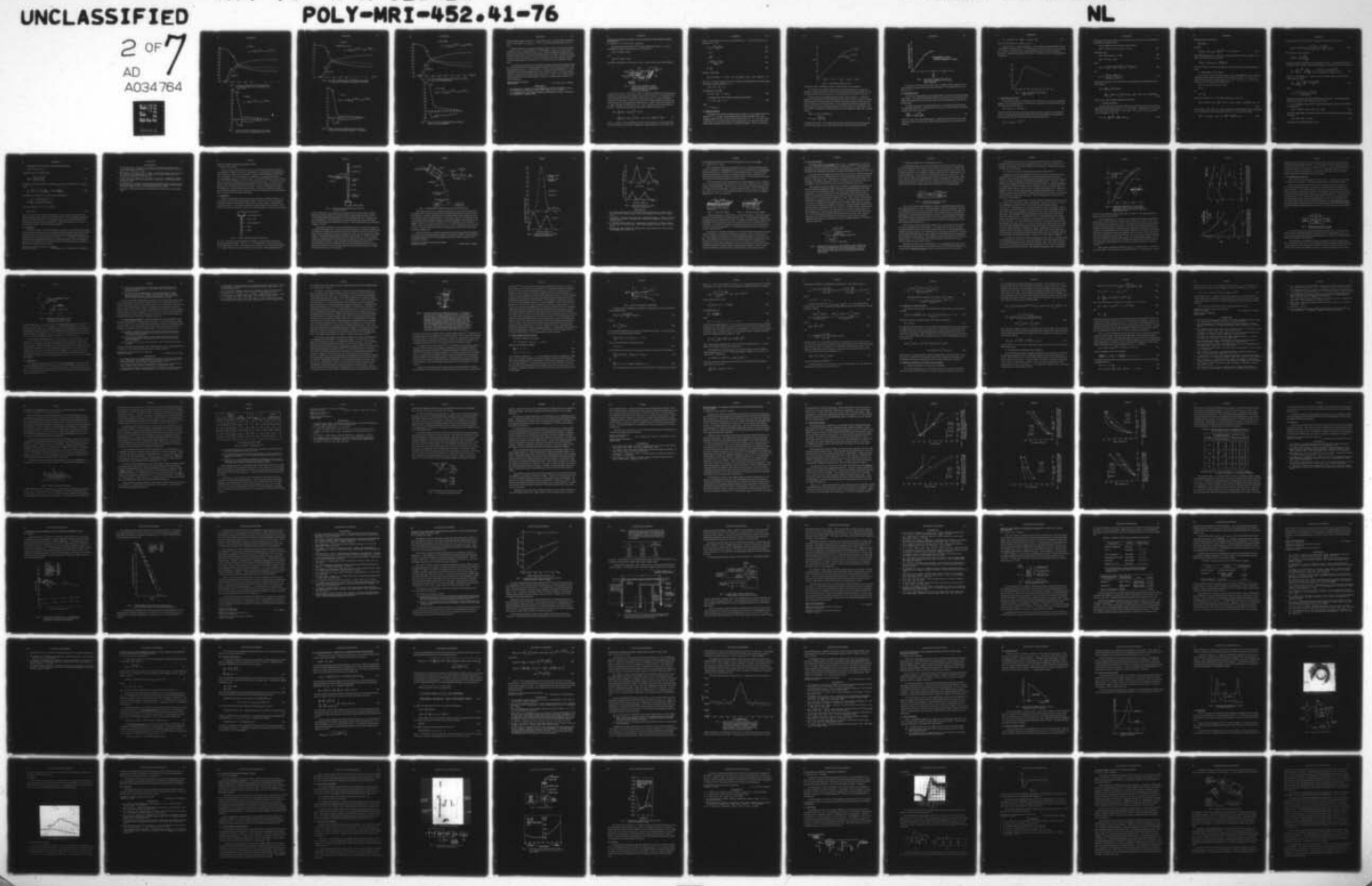
C. Numerical Results for a Crack Originating at the Surface

Computation of R, T and the scattered power $1 - |R|^2 - |T|^2$ were made for slip cracks originating at the surface of aluminum and polyethelene. For comparison both choices in Eqs. (16) and (17) of f(z) were used for the aluminum solid.

Figure 2 shows the variation of $|R|$, $|T|$ and $1 - |R|^2 - |T|^2$ as a function of crack depth h normalized to the Rayleigh wave length $\lambda_R = 2\pi/k_R$. In computing the results presented in Fig. 2, Eq. (16) was used for f(z). For $h/\lambda_R < 0.2$, the scattering of the Rayleigh wave is very small, while for h/λ_R in the range 0.2 to 0.6, the scattering increases sharply. As h/λ_R increases beyond 0.6, $|R|$ approaches 0.68; $|T|$ approaches 0.32 and the fraction of the incident power scattered into bulk waves approaches 0.43. For $h > \lambda_R$, the scattering is not sensitive to crack depth since most of the power in the incident surface wave is concentrated near the surface. In Fig. 3 we have

AD-A034 764 POLYTECHNIC INST OF NEW YORK BROOKLYN MICROWAVE RESE--ETC F/G 20/3
PROGRESS REPORT NUMBER 41, 15 SEPTEMBER 1975 THROUGH 14 SEPTEMB--ETC(U)
NOV 76 A A OLINER F44620-74-C-0056
UNCLASSIFIED POLY-MRI-452.41-76 NL

2 OF 7
AD
A034764



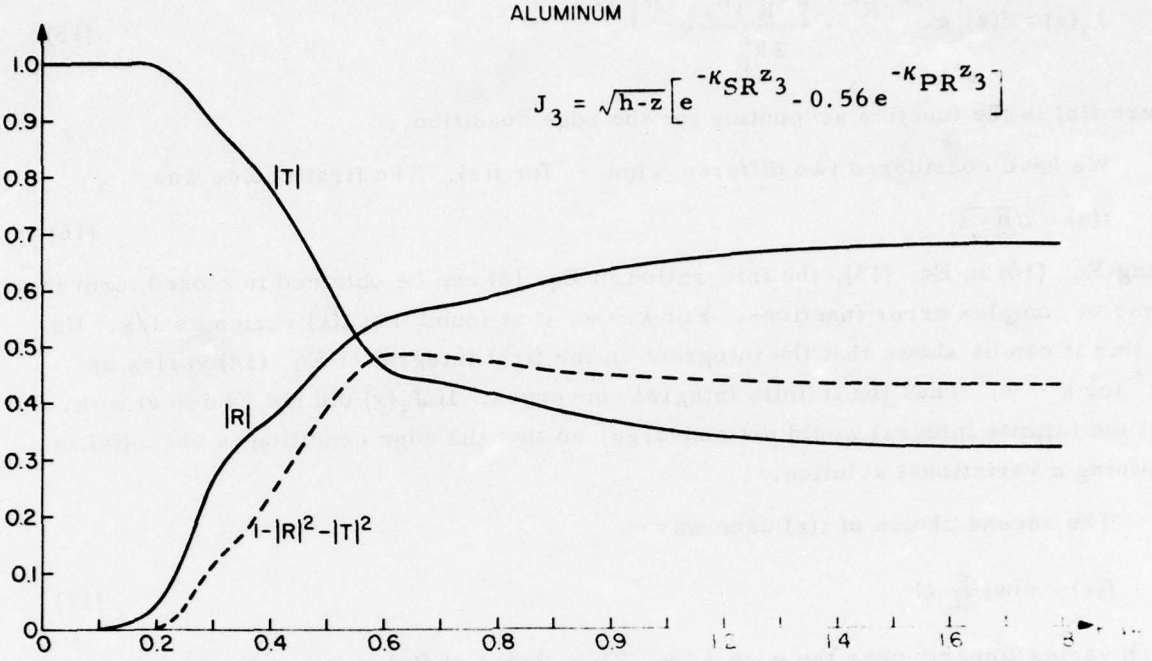


Fig. 2. R, T and scattered power for Rayleigh wave incident on a slip crack in aluminum using $f(z) = \sqrt{h - z}$.

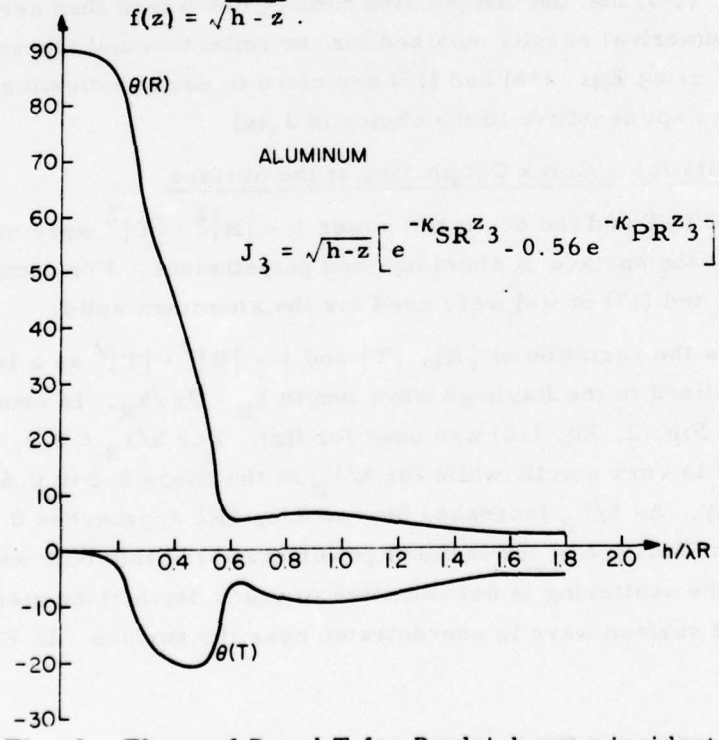


Fig. 3. Phase of R and T for Rayleigh wave incident on a slip crack in aluminum using $f(z) = \sqrt{h - z}$.

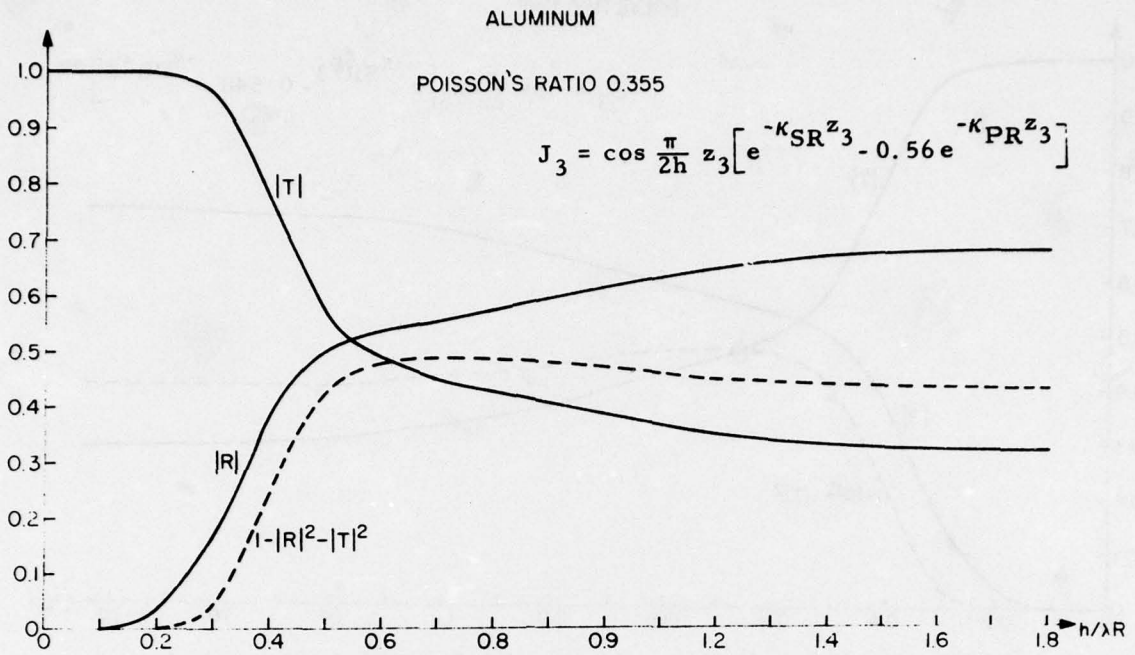


Fig. 4. R, T and scattered power for Rayleigh wave incident on a slip crack in aluminum using f(z) = cos(π z/2h).

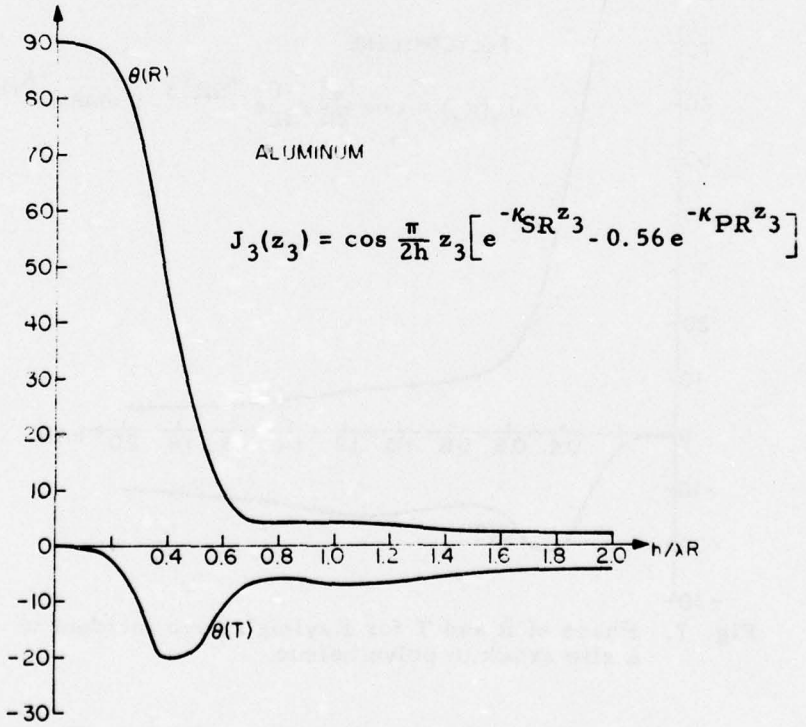


Fig. 5. Phase of R and T for Rayleigh wave incident on a slip crack in aluminum using f(z) = cos(π z/2h).

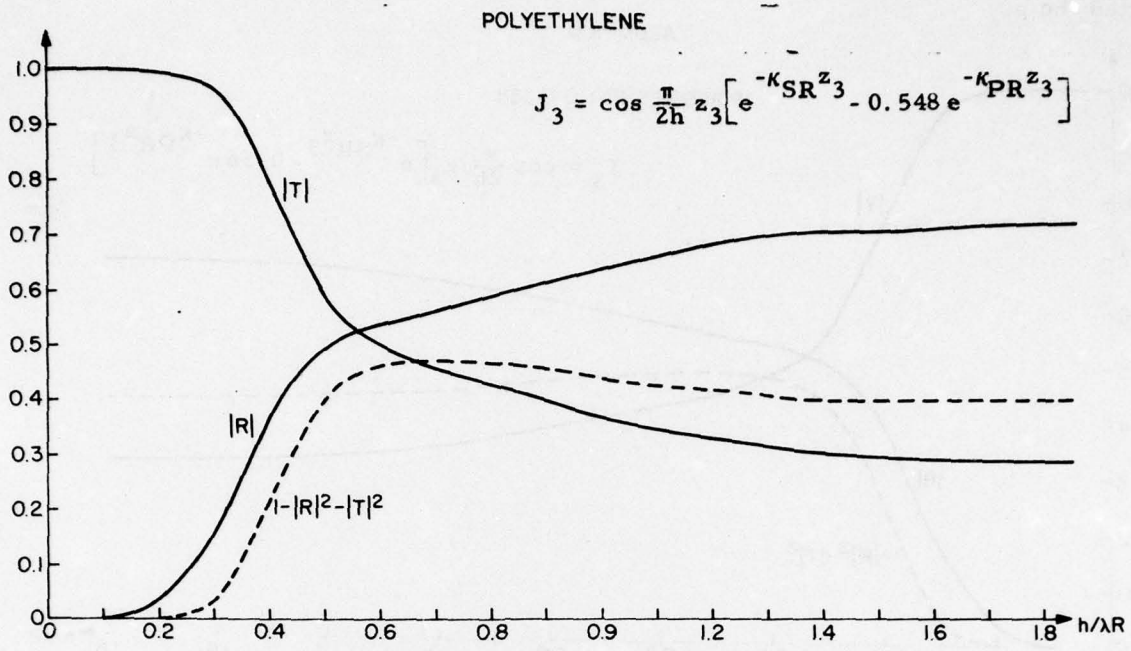


Fig. 6. R, T and scattered power for Rayleigh wave incident on a clip crack in polyethelene.

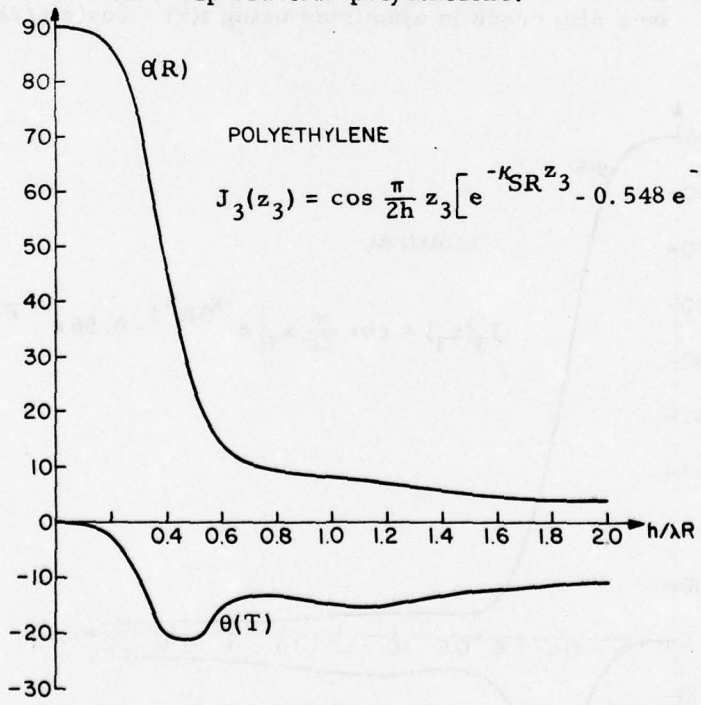


Fig. 7. Phase of R and T for Rayleigh wave incident on a slip crack in polyethelene.

plotted the phase angles of R and T. As with $|R|$ and $|T|$, rapid variation in the phase occurs in the range $0.2 < h/\lambda_R < 0.6$. For $h > \lambda_R$, the phase angles approach a constant.

Results obtained for aluminum but using the choice in Eq. (17) for f(z) are plotted in Figures 4 and 5. Comparing Figs. 2 and 4 and Figs. 3 and 5 it is seen that the results are not sensitive to f(z), and hence $J_3(z)$. The major difference appears to be in the phase angles of R and T, and even these differences are small. Because the computing time is much less when Eq. (17) is used, and since the results using Eqs. (16) and (17) are nearly identical, the form of Eq. (17) for f(z) is used for subsequent calculations.

It is desirable to know how the scattering will depend on the material being studied. It can be shown that the variational expression depends on the material properties only through Poisson's ratio σ. For aluminum $\sigma = 0.355$, while for polyethylene $\sigma = 0.458$. Computations made for polyethylene are shown in Figures 6 and 7. It is seen that this change in σ produces only minor changes in R, T and $1 - |R|^2 - |T|^2$. Further computations will be made over a wider range of values of σ to determine the influence of σ on the scattering.

National Science Foundation
ENG75-00569

Y. L. Hou and H. L. Bertoni

REFERENCES

1. G. C. Sih and J. F. Loeber, "Wave Propagation in an Elastic Solid With a Line of Discontinuity or Finite Crack," Quart. Appl. Math., 27, 193- (1969).
2. N. I. Muskhelishvili, "Some Basic Problems of Mathematical Theory of Elasticity," P. Noordhoff Co., Holland (1953).

STABILITY AND NOISE ANALYSIS OF THE ACOUSTOELECTRIC PHASE-LOCKED LOOP DEVICE

H. Schachter, F.A. Cassara and L. Rosenheck

We have previously reported[1, 2, 3] that the configuration of Fig. 1 acts as an acoustoelectric phase-locked loop (PLL). In summary if

$$e_1(t) = A \sin(W_1 t + \psi_1(t))$$

and

$$e_2(t) = B \cos(W_1 t + \psi_2(t))$$

are added at the input to the transducer of Fig. 1, they will give rise to two Rayleigh

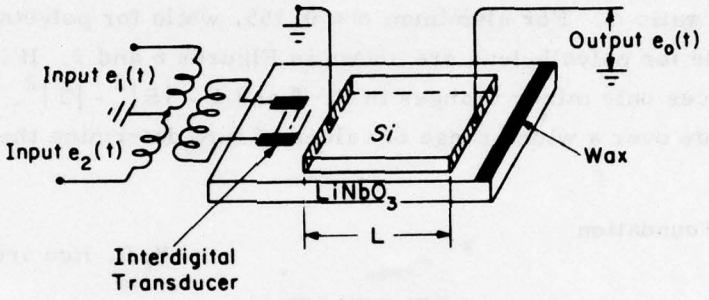


Fig. 1. Structure of a nonlinear acoustic surface-wave device. It performs both the functions of multiplication and low-pass filtering.

acoustic surface waves. The electric fields carried by the Rayleigh waves will generate oscillatory charges and electric fields in the nearby silicon plate. The nonlinear interaction between these charges and electric fields gives rise to currents at difference frequencies in the silicon. Finally, tne open circuit voltage resulting from integrating these currents over the length of the semiconductor is used to control the frequency of a voltage controlled oscillator (VCO), i.e., $\dot{\psi}_2(t) = \beta e_0(t)$ where β is the sensitivity of the VCO. The relation governing the behavior of the PLL was shown to be

$$\dot{\psi}_2(t) = \frac{\beta}{2n_0}(E_1 n_1 + E_2 n_2) \int_0^L U(t - \tfrac{z}{v})\, dz$$

$$+ \frac{\beta}{2n_0}(E_1 n_2 + E_2 n_1) \int_0^L \sin[\omega_1 t - \omega_2 t + \psi_1(t) - \psi_2(t)]\, U(t - \tfrac{z}{v})\, dz \qquad (1)$$

where: E_1 and E_2 are the amplitudes of the two electric fields, n_1, n_2 the respective generated charges, n_0 the thermal equilibrium charge distribution in the silicon, and

finally, v the propagation velocity of the Rayleigh waves. It was found useful to normalize Eq. (1). Let

$$\omega_n^2 = \beta v \frac{E_1 n_2 + E_2 n_1}{2n_0} \tag{2a}$$

$$\tau = \omega_n t \tag{2b}$$

$$a = \frac{\omega_n L}{v} \tag{2c}$$

$$\gamma = \frac{\beta L(E_1 n_1 + E_2 n_2)}{2n_0 \omega_n} \tag{2d}$$

$$W = \frac{\omega}{\omega_n} \tag{2e}$$

then Eq. (1) becomes

$$\dot{\psi}_2(\tau) = \frac{\gamma}{a}(\tau U(\tau) - (\tau - a)\, U(\tau - a)) + \int_{\tau-a}^{\tau} \sin(W_1\lambda - W_2\lambda + \psi_1(\lambda) - \psi_2(\lambda))\, d\lambda \tag{3}$$

Let $\dot{\psi}_1(\tau) = 0$, then in steady state the VCO frequency will follow that of the input signal with a phase constant difference, i.e.,

$$\psi_2(\tau) = W_1\tau - W_2\tau - \psi_s$$

and solving Eq. (3) yields

$$\Delta W = W_1 - W_2 = \gamma + a \sin \psi_s \tag{4a}$$

According to Eq. (4a) the hold-in range will vary between

$$\gamma - a < W_1 - W_2 < \gamma + a \tag{4b}$$

since

$$-1 < \sin \psi_s < 1$$

A. Stability Analysis

As shown in Eq. (4b) the maximum hold-in range (i.e., lock-range) for the PLL should be $\gamma + a$. In order to verify dynamically the static prediction of Eq. (4b), computer solutions of Eq. (3) were carried out for values of $\gamma = 0, a, 2a$ with a and ψ_s as parameters. For a given value of a the maximum of ψ_s for which the solution was stable were calculated and the results are plotted in Figure 2. Since γ/a was kept a

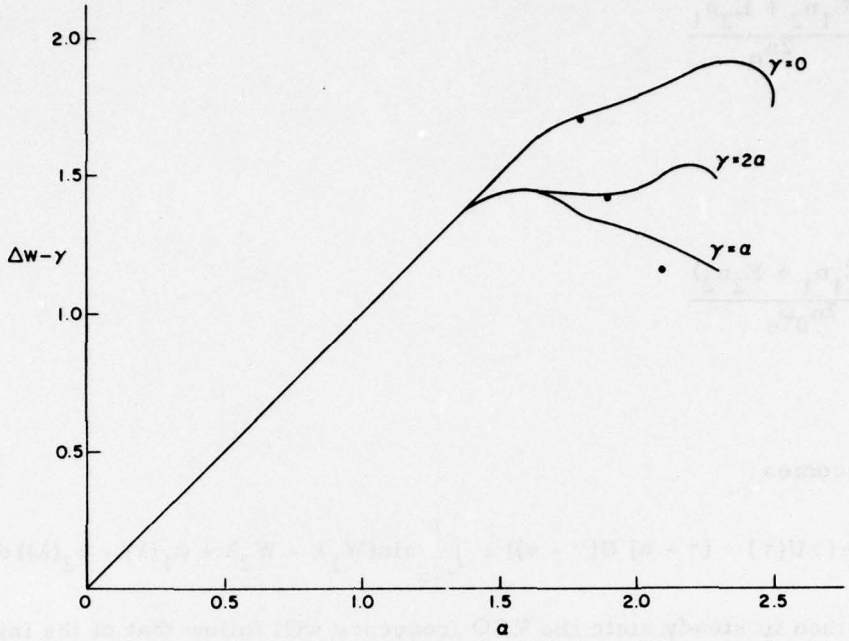


Fig. 2. PLL hold-in-range versus α.

constant for these calculations, only the variations with α are considered.

For α small, the maximum hold-in range varies linearly with α as predicted. However, for large values of α, the inherent nonlinearity of the system limits the hold-in range, and the value of ΔW remains a constant with an increase in the value of α.

These results were verified experimentally as shown in Figure 3. The experimental set-up was identical to that described in References 1 and 2. Here the maximum hold-in range versus the variations on A/B (the ratio of the amplitude of the input voltage to the output of the VCO) is plotted. In this experiment the output amplitude of the VCO was kept constant and the DC component of the VCO input was compensated. Analytically this corresponds to setting the first term on the right hand side of Eq. (3) equal to zero.

With $\gamma = 0$, Eq. (4) reduces to

$$\Delta\omega = \omega_n \alpha = \frac{\beta L E_2 n_2}{n_0} \frac{A}{B} \tag{5}$$

A comparison of Figs. 2 and 3 shows that the hold-in range increases linearly but saturates for large values of α or equivalently (see Eq. (5)) for large values of A/B.

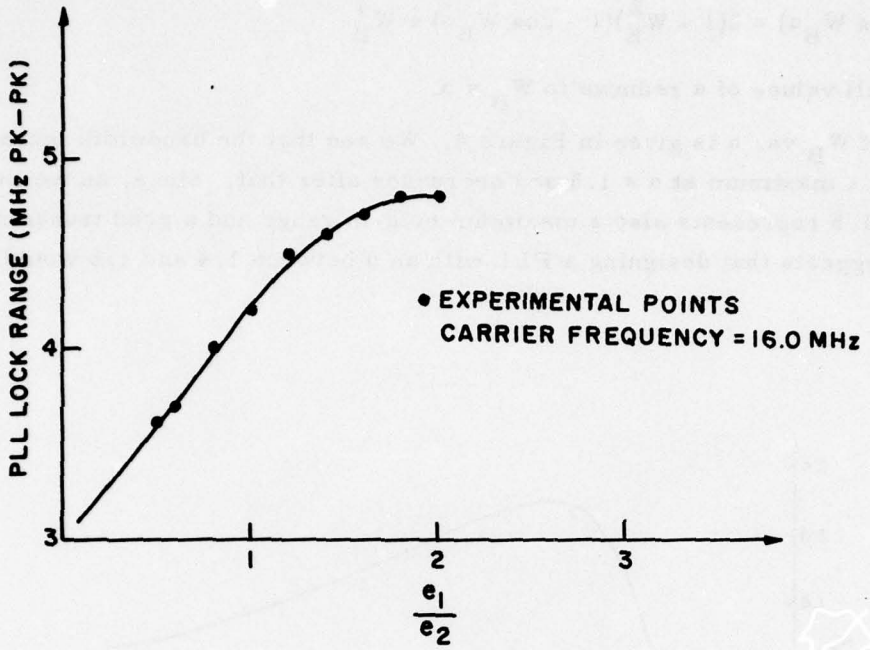


Fig. 3. PLL hold-in-range versus ratio of input signal amplitude to VCO output signal amplitude.

In summary we observe that in order to get a maximum hold-in range and at the same time have a system with a good transient behavior[1,3] we should choose an a of about 1.4 to 1.6.

B. Bandwidth of PLL

The PLL is accompanied by an IF bandpass filter in the input and a low pass filter in the output. The closed loop bandwidth of the PLL is optimal at about twice the bandwidth of the IF filter and of course much wider than the bandwidth $2f_m$ of the post detection filter.[4]

In order to design the accompanying filters it is of interest to know the closed loop bandwidth of the PLL. Linearizing Eq. (3) it is easy to show that the closed loop response of the PLL is given by

$$\frac{\psi_2(W)}{\psi_1(W)} = \frac{1 - e^{-jWa}}{1 - W^2 - e^{-jWa}} \tag{6}$$

where $W = \omega/\omega_n$ is the normalized frequency. Note that the PLL is a low pass filter whose gain is one at DC and whose 3 dB bandwidth W_B is given by the solution to the equation:

$$4(1 - \cos W_B\alpha) = 2(1 - W_B^2)(1 - \cos W_B\alpha) + W_B^4 \tag{7}$$

which for small values of α reduces to $W_B = \alpha$.

A plot of W_B vs. α is given in Figure 4. We see that the bandwidth increases with α and reaches a maximum at $\alpha = 1.5$ and decreases after that. Since, as we saw earlier, an α of about 1.5 represents also a maximum hold-in range and a good transient reponse, this suggests that designing a PLL with an α between 1.4 and 1.6 would represent an optimum.

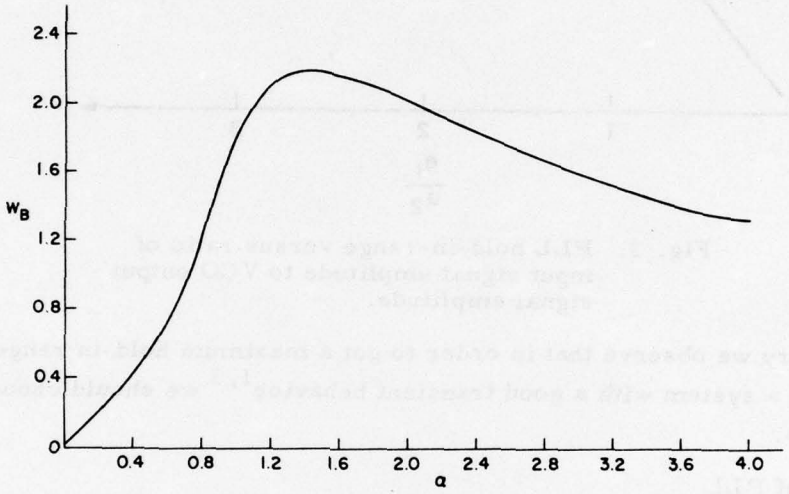


Fig. 4. PLL normalized 3 dB bandwidth (W_B) versus α.

C. Noise Response of PLL

When the input signal is corrupted by noise, additional a.m. and f.m. modulation appears. Let us assume that the corrupting noise is gaussian with zero expected value. Then the noise can be written as

$$n(t) = a(t) \cos \omega_1 t - b(t) \sin \omega_1 t \tag{8}$$

where a(t) and b(t) are slowly varying gaussian stochastic processes of zero expected value. Assuming that the dominant pole of the low pass equivalent of the predetection IF filter is ω_{IF}, then the autocorrelation function of a(t) and b(t) is

$$R_{aa}(\tau) = R_{bb}(\tau) = \sigma^2 e^{-\omega_{IF}\tau}$$

where ω_{IF} is the center frequency of the input signal and $2\omega_{IF}$ is the 3 dB bandwidth of the IF filter preceding the PLL.

With the addition of noise the input signal becomes

$$e_1(t) = A \sin(\omega_1 t + \psi_1) + a \cos \omega_1 t - b \sin \omega_1 t \tag{9a}$$

Rewriting we get

$$e_1(t) = AR(t) \sin(\omega_1 t + \psi_1 + \theta(t)) \tag{9b}$$

$$e_2(t) = B \cos(\omega_2 t + \psi_2(t)) \tag{9c}$$

where

$$R(t) = \sqrt{1 + (\frac{a}{A})^2 + (\frac{b}{A})^2 + 2\frac{a}{A}\sin\psi_1 - 2\frac{b}{A}\cos\psi_1} \tag{9d}$$

and

$$\tan\theta = \frac{\frac{a}{A}\cos\psi_1 + \frac{b}{A}\sin\psi_1}{1 + \frac{a}{A}\sin\psi_1 - \frac{b}{A}\cos\psi_1} \tag{9e}$$

With $e_1(t)$ and $e_2(t)$ as in Eqs. (9b) and (9c) the normalized governing relation of the PLL becomes

$$\dot{\psi}_2(\tau) = \frac{A}{2B}\int_{\tau-a}^{\tau} R^2(\lambda)\, U(\lambda)\, d\lambda + \frac{B}{2A}\int_{\tau-a}^{\tau} U(\lambda)\, d\lambda + \int_{\tau-a}^{\tau} R(\lambda) \sin[(W_1 - W_2)\lambda + \psi_1(\lambda) + \theta(\lambda) - \psi_2(\lambda)]\, U(\lambda)\, d\lambda \tag{10}$$

where τ and a were defined in Equations (2b) and (2c).

1. Rectification Effects

We notice that the noise will contribute some rectification which will result in a shift in the hold-in range of the PLL. For small noise, assuming that in steady state the average value of the $\sin[(W_1 - W_2)\lambda + \psi_1(\lambda) + \theta(\lambda) - \psi_2(\lambda)] = \sin\psi_s$ we get with $\psi_1(\tau) = 0$ and $\dot{\psi}_2(\tau) = W_1 - W_2$

$$W_1 - W_2 = \frac{A}{2B}\, a\, \overline{R^2} + a\frac{B}{2A} + a\,\overline{R}\sin\phi_s \tag{11a}$$

Realigning and recognizing that

$$\frac{A}{2B} + \frac{B}{2A} = \frac{\gamma}{a}$$

we obtain

$$\Delta W_{\text{noise}} = \Delta W_{\text{no noise}} + \frac{A}{2B} a(\overline{R^2} - 1) + a \sin \phi_s(\bar{R} - 1) \tag{11b}$$

where $\overline{R^2}$ and $\bar{R}$ represent, respectively, the expected values of $R^2(\tau)$ and $R(\tau)$. Carrying out the expectations we find that

$$\Delta W_{\text{noise}} - \Delta W_{\text{no noise}} = a \frac{\sigma^2}{A^2}\left(\frac{A}{B} + \frac{3}{2}\right)$$

This shows that the rectification brought by noise will increase the maximum hold-in range.

2. Output Signal-to-Noise Ratio

We will now calculate the output signal-to-noise ratio (SNR) of the PLL under the conditions of input carrier mixed with a weak gaussian noise ($\frac{\sigma}{A} << 1$). That is, we assume that the system is at equilibrium with $\psi_2 = (W_1 - W_2)\tau + \psi_s$, $\psi_1 = 0$ and now we introduce noise which disturbs this equilibrium.

Since we consider weak noise, we let

$$R(\tau) \approx 1$$

and

$$\theta(\tau) \approx \frac{a}{A}$$

under the assumption of $\tau > 0$ and differentiating Eq. (10) we get

$$\ddot{\psi}_2(\tau) = \sin\left[(W_1 - W_2)\tau + \frac{a(\tau)}{A} - \psi_2(\tau)\right] - \sin\left((W_1 - W_2)(\tau - a) + \frac{a(\tau - a)}{A} - \psi_2(\tau - a)\right) \tag{12a}$$

If we let $\psi_2(\tau) = (W_1 - W_2)\tau + \psi_s + \phi_2(\tau)$ where $\phi_2(\tau)$ is the disturbance from equilibrium of $\psi_2(\tau)$ introduced by noise and assume that $a(\tau)$ $A - \phi_2(\tau)$ is small we get:

$$\ddot{\phi}_2(\tau) + \cos \psi_s(\phi_2(\tau) - \phi_2(\tau - a)) = \left(\frac{a(\tau)}{A} - \frac{a(\tau - a)}{A}\right) \cos \psi_s \tag{12b}$$

Let $S_{a/A}(W)$ be the spectrum of a/A and $S_{\dot{\phi}_2}(W)$ the spectrum of $\dot{\phi}_2(\tau)$ then

$$S_{\dot{\phi}_2}(W) = S_{a/A}(W) \frac{2 \cos^2 \psi_s (1 - \cos W a) W^2}{W^4 + 2 \cos \psi_s (\cos W a - 1)(W^2 - \cos \psi_s)} \tag{13}$$

The spectrum of a/A is given by

$$S_{a/A}(W) = \frac{\sigma^2}{A^2} \frac{2 W_{IF}}{W^2 + W_{IF}^2}$$

when W_{IF} is half the bandwidth of the predetection IF filter. Assuming that the PLL is followed by a post detection filter of bandwidth W_M the variance of $\dot{\phi}_2(\tau)$ will be given by

$$\sigma^2_{\dot{\phi}_2} = \frac{\sigma^2 2}{A^2 2\pi} \int_0^{W_M} \frac{2 W_{IF}}{W^2 + W_{IF}^2} \cdot \frac{2 \cos^2 \psi_s (1 - \cos W a) W^2 dW}{W^4 + 2 \cos \psi_s (\cos W a - 1)(W^2 - \cos \psi_s)} \tag{14}$$

if $W_M a \ll 1$, and $W_M / W_{IF} \ll 1$, which is usual in the case of PLL, Eq. (14) can be easily integrated to give

$$\sigma^2_{\dot{\phi}_2} = \frac{\sigma^2}{\pi A^2 W_{IF}} \left(\frac{2 W_M}{D^2} - \frac{2 \arctan W_M D}{D^3} \right) \tag{15}$$

where

$$D^2 = \frac{12 - 12 a^2 \cos \psi_s - a^4 \cos^2 \psi_s}{12 a^2 \cos^2 \psi_s}$$

Equation (15) represents the detected RMS output noise power. The detected output signal power under no noise conditions is $[W_1 - W_2]^2$.

In a conventional first order PLL the governing equation with a weak corrupting gaussian noise is[5]

$$\dot{\psi}_2(t) = C \sin((W_1 - W_2)t + \psi_1(t) + \theta(t) - \psi_2(t)) \tag{16}$$

where the notation used is consistent with the acoustic PLL notation previously defined. As before,

$$\psi_2(t) = (W_1 - W_2)t + \psi_s - \phi_2(t) \tag{17}$$

where $\phi_2(t)$ is the disturbance due to noise.

Substituting Eq. (17) into Eq. (16) and linearizing we get

$$\dot{\phi}_2(t) = C(\theta - \phi_2) \cos \psi_s \tag{18}$$

Taking the spectrum of $\dot{\phi}_2(t)$ we find

$$S_{\dot{\phi}_2}(\omega) = \frac{S_\theta C^2 \cos^2 \psi_s \omega^2}{\omega^2 + C^2 \cos^2 \psi_s} \tag{19}$$

Assuming a post detection filter of bandwidth ω_M and S_θ a constant over $0 < \omega < \omega_M$ we obtain

$$\sigma^2_{\dot{\phi}_2} = \frac{S_\theta}{\pi} C^3 \cos^3 \psi_s \left[\frac{\omega_M}{C \cos \psi_s} - \arctan \frac{\omega_M}{C \cos \psi_s} \right] \tag{20}$$

Comparing, we see that $C \cos \psi_s$ in Eq. (20) corresponds to

$$\frac{1}{D} = \sqrt{\frac{12 a^2 \cos^2 \psi_s}{12 - 12 a^2 \cos \psi_s - a^4 \cos^2 \psi_s}}$$

in Eq. (15) which for $a \cos \psi_s << 1$ becomes

$$\frac{1}{D} \simeq a \cos \psi_s \tag{21}$$

That is for small a the acoustic PLL behaves as a first order PLL with closed loop bandwidth a which agrees with previous interpretations. Also, substitution of Eq. (21) in Eq. (15) reveals that the r.m.s. output noise $\sigma_{\dot{\phi}_2}$ for the conventional first order PLL becomes identical with the acoustic PLL. The noise response for the PLL outside the linear region is currently being investigated.

D. Conclusions

The hold-in-range, closed loop bandwidth, and weak noise response of the acoustoelectric PLL have been derived. The computer studies presented reveal that a practical "rule-of-thumb" to achieve good transient response and maximum hold-in-range is to design the PLL's normalized "delay time" a equal to 1.5. In addition, it has been demonstrated that the detected output signal-to-noise ratio response of the acoustoelectric PLL and the conventional PLL are identical when the input FM carrier is corrupted by additive weak narrowband gaussian noise.

Joint Services Technical Advisory Committee
F44620-74-C-0056 H. Schachter, F.A. Cassara and L. Rosenheck

REFERENCES

1. H. Schachter, W.C. Wang and F.A. Cassara, "An Integrated Acoustic Phase-Locked Loop Device," IEEE International Electron Devices Meeting Technical Digest, 159-162 (December 1974).

2. K.S. Meng, H. Schachter and W.C. Wang, "A New Application of Acoustic Surface Wave Devices: The Acousto-Electric Phase-Locked Loop," IEEE Trans. on Ultrasonics Symposium (November 1974).

3. H. Schachter, K.S. Meng, W.C. Wang and F.A. Cassara, "Application of Surface Acoustic Wave Devices to Communications," IEEE Trans. on Electron Devices (September 1976).

4. M. Unkauf and R. Schulman, "Experimental Signal/Noise-Ratio Comparison of the Second Order Phase-Locked Loop and the Second Order Frequency-Locked Loop," Electronics Letters, 4, No. 26, 585-586 (December 27, 1968).

5. A. Viterbi, "Principles of Coherent Communication," (New York: McGraw Hill, 1966).

SPECULAR REFLECTION FROM SODIUM VAPOR

V. Nangia and B. Senitzky

Use of specular reflection from vapors as a means for developing optical filters was suggested a couple of years ago.[1] Investigations, both theoretical and experimental, were carried out on mercury vapor filters[2,3] and the results obtained at 2537°A were very encouraging. The mercury vapor cell operated as a narrowband filter with a 0.14A bandwidth, having a 7° acceptance angle. It provided 28 dB peak-to-skirt rejection ratio with a peak 14% reflectance on a one inch diameter aperture. This filter was used to get a monochromatic photograph of distribution of mercury atoms in a flame and to determine the concentration of mercury in water. A list of 26 elements that could be investigated for a similar purpose was proposed.[1] In order to investigate the application of the device principle to one of the more "difficult" elements on the aforementioned list, we decided to build a sodium vapor filter.

A. Experiment

Sodium at high temperatures attacks quartz; therefore we decided to build a cell consisting of a nickel housing with sapphire window. Attempts made to bond sapphire to nickel to withstand temperatures around 900°C were unsuccessful. Consequently, it was decided to use a commercially available seal that withstands temperatures to about 600°C. This seal was attached to one end of a nickel tubing (Fig. 1) by electro-deposition

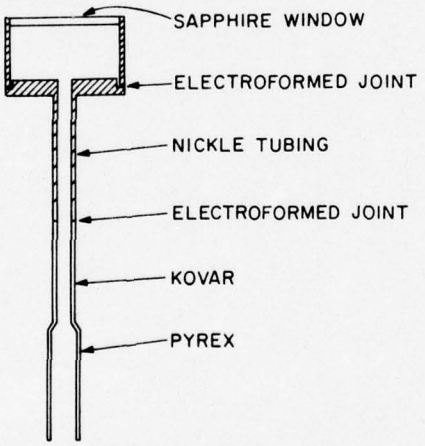


Fig. 1. Vapor cell joined to Kovar-Pyrex seal.

of nickel. The other end of the tubing was attached to a Pyrex-Kovar seal (also by electro-deposition of nickel). These electro-formed joints have been found to be vacuum-tight during heat-cycling. This assembly was connected to a vacuum station (Figure 2). A sodium ampule was connected directly above the cell. After baking the cell at

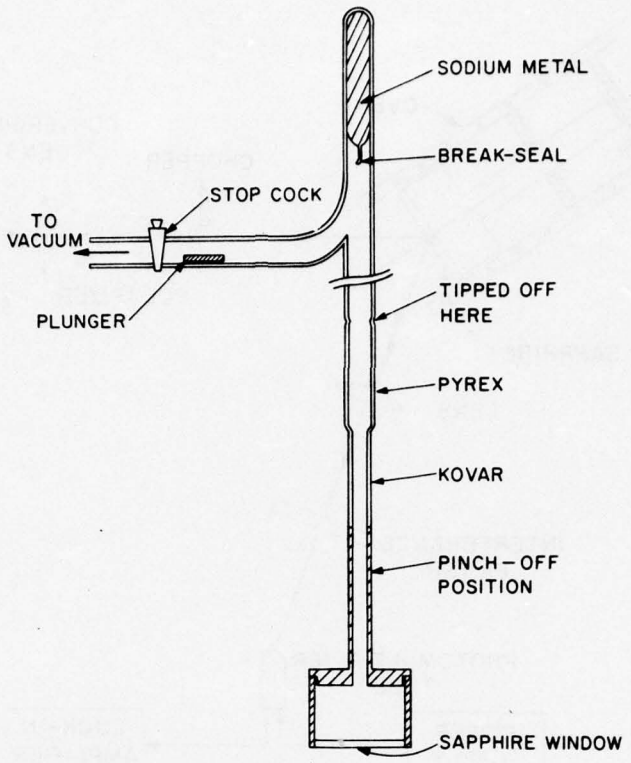


Fig. 2. Vapor cell and sodium ampule connected to vacuum station.

600°C, sodium metal was melted with a propane torch (care being taken for uniform heating all along the length of the ampule and the connecting glass tubing), so that a small droplet gradually flowed down into the nickel tubing of the cell. The glass tube was then tipped off. The nickel tubing was, thereafter flattened and pinched using a pinch-off tool which exerted 4500 psi of pressure. The pinched edge was electron-beam welded in vacuum.[4]

The cell was mounted in an oven and oriented to receive, at Brewster's angle, incoming polarized light from a narrow-slit object illuminated by a sodium discharge lamp. The intensities of light reflected both from the front surface and the back surface of the window were measured using a photo-multiplier tube with a lock-in amplifier and a light chopper. The experimental arrangement used is shown in Figure 3. The two images produced on reflection from the two surfaces were separated so that each could be scanned by the photo-multiplier. The birefringence of the sapphire did not affect our measurements because we oriented the optic axis of the window perpendicular to the direction of polarization of the incoming signal (we determined this position by finding the intensity minima of the two images).

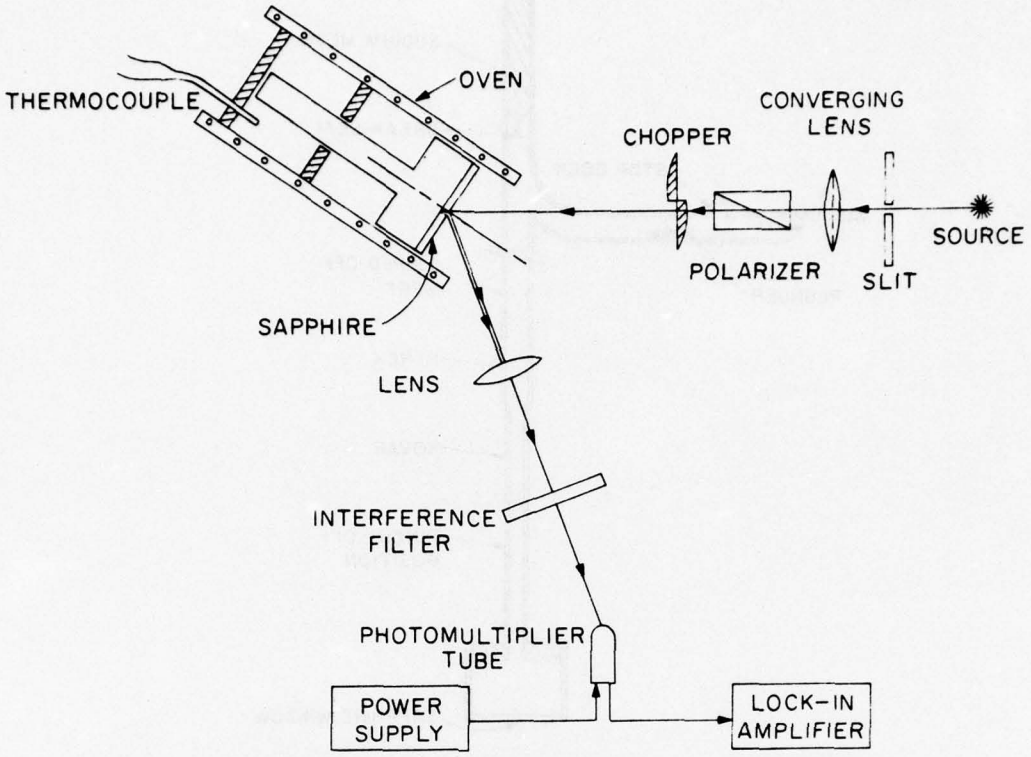


Fig. 3. Experimental arrangement for intensity measurements.

The results obtained for the reflected intensities for two different temperatures are shown; these curves show the variation of reflected intensity as a function of position (Figure 4). Similar experiments were performed using a He-discharge lamp and the results for that are shown in Figure 5. From these curves, it is found that, in the case of sodium-line the reflected intensity from vapor-sapphire surface relative to air-sapphire surface, increases by a factor of 11.7 for temperature rise from 25°C to 595°C. In the case of sodium-line, this factor is 1.6 for a temperature rise from 25°C to 585°C. Thus the relative reflectance for sodium-line is about seven times larger than that for helium line.

The above results are conclusive proof that we are obtaining specular reflection from sodium vapor and indicate that a sodium vapor filter is feasible. We are presently obtaining quantitative data on this phenomenon, which we hope to compare to our theroetical estimates.

Joint Services Technical Advisory Committee
F44620-74-C-0056

V. Nangia and B. Senitzky

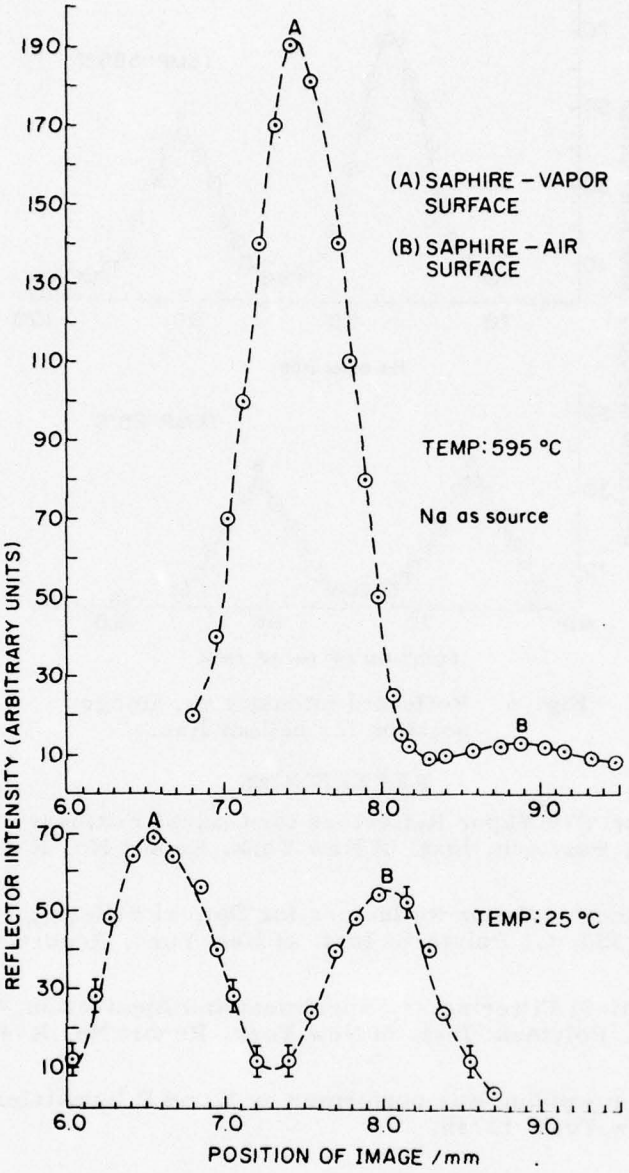


Fig. 4. Reflected intensity vs. image position for sodium line.

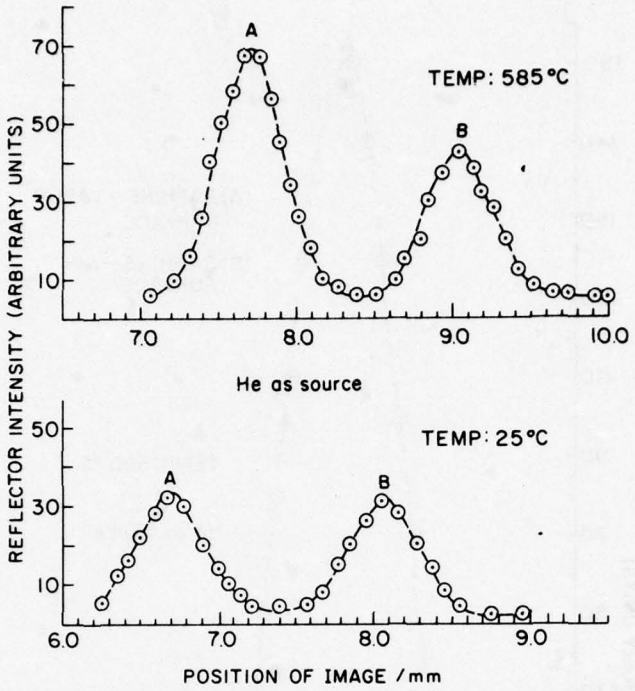


Fig. 5. Reflected intensity vs. image position for helium line.

REFERENCES

1. B. Senitzky, "Selective Vapor Reflectors for Optical Filtering," Progress Report No. 37 to JSTAC, Polytech. Inst. of New York, Report No. R-452.37-72, 88-94 (1972).

2. B. Senitzky, "Selective Vapor Reflectors for Optical Filtering -- Theory," Progress Report No. 38 to JSTAC, Polytech. Inst. of New York, Report No. R-452.38-73, 139-143 (1973).

3. B. Senitzky, "Optical Filtering -- Experiment and Application," Progress Report No. 38 to JSTAC, Polytech. Inst. of New York, Report No. R-452.38-73, 134-139 (1973).

4. The electron-beam welding was performed by E and B Industries, 81 Gazza Blvd., Farmingdale, New York 11735.

WAVEGUIDES FOR INTEGRATED OPTICS FORMED BY METAL PLATINGS

A.A. Oliner and S.T. Peng

Most waveguides which have been proposed for use in integrated (planar) optics consist of combinations of dielectric materials. It is less well known that metal strips plated on thin films on a substrate can also guide optical energy. Such metal strips have been incorporated into devices for integrated optics (such as modulators and switches) which require the application of electric fields, but the developers have often been unaware of the guiding properties of those devices.

The two structures to be discussed here are shown in Figures 1 and 2. Figure 1 shows an optical slot waveguide, formed of a slot or gap of unplated film on a substrate between two plated regions. The discussions in the literature for this waveguide type are incorrect for one polarization, as will be explained below. The second configuration, shown in Fig. 2, is a new waveguiding structure consisting simply of a strip of metal plated on a film on a substrate.

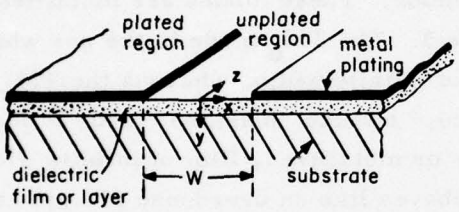


Fig. 1. The slot waveguide.

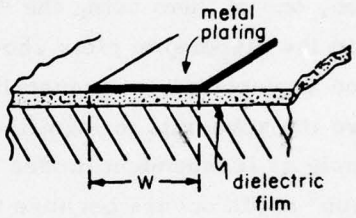


Fig. 2. The metal-strip waveguide.

At optical frequencies, metals behave as overdense plasmas, so that they possess dielectric constants that are essentially negative real. It is this property that permits the guiding to occur without excessive loss. In addition, the performance of these guides is different from that of guides at microwave frequencies that may resemble them superficially. For example, the optical slot guide discussed here behaves quite differently from the microwave slot guide; it is actually an analogue of the slot waveguide for acoustic surface waves.[1,2]

The method of calculation employed by us in determining the propagation properties of these structures is the transverse resonance procedure, which is widely used at microwave frequencies and which was in fact originally developed in a microwave context. This method can provide more accurate results for optical waveguides than other procedures which have so far been used, and yet it is simple and straightforward. Furthermore, the procedure readily takes into account the presence of geometrical discontinuities between the various regions comprising the waveguide structure, and is capable of systematic improvement in successive steps, to the accuracy desired.

A. The Slot Waveguide

The slotted metal-clad waveguide shown in Fig. 1, designated here as the "slot waveguide," has been considered previously in the literature.[1,3,4] This waveguiding structure was first proposed and analyzed by one of the present authors,[1] and it was soon afterwards independently proposed, measured and analyzed by Japanese workers.[3,4]

The modes which propagate may be divided into two classes: those for which only an electric field, or only a magnetic field, is present in the vertical direction, perpendicular to the dielectric film. These modes correspond to the E^y and E^x modes, respectively, described by Marcatili.[5] We are concerned now with the former class, or polarization, since it is for this class that previous analyses are incorrect.

The transverse resonance approach to the solution of waveguiding problems proceeds by viewing the guiding structure transversely and identifying the modes which exist in each of the constituent regions. For the E-vertical polarization, it is well known that, at the lower frequency end of the propagation range, only one mode is present in the unplated region whereas two modes are present simultaneously in the plated region, one of them being the "plasmon"[6] mode. These modes are identified as solid lines in the dispersion plots shown in Figure 3. The TM_u mode is the one which is present on an unplated, or unmetallized, film of infinite width, whereas the TM_m mode, where the subscript m signifies "metallized," and the "plasmon" mode P occur simultaneously as independent modes on a plated, or metallized, film of infinite width. The "plasmon" mode occurs because the metal behaves like an overdense plasma at optical frequencies, and this mode would exist even if the dielectric film were absent. It is, in fact, the dominant mode for this polarization (E vertical), and it cannot be neglected; for H-vertical polarization, the "plasmon" mode is absent. It is this "plasmon" mode which was omitted from previously-published analyses[3,4] for the E-vertical polarization.

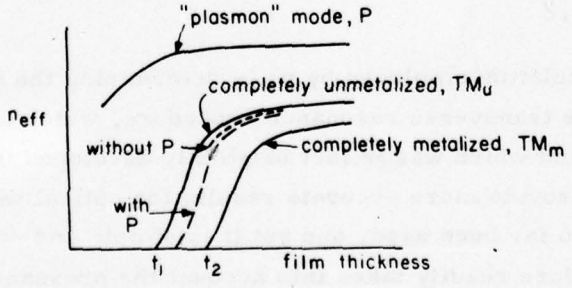


Fig. 3. Dispersion curves (n_{eff} vs. film thickness), shown dashed, for the lowest E-vertical mode in the slot waveguide. Curves for the constituent modes are shown solid (the vertical spacing between the P and the TM_u curves is generally greater than is shown here).

1. Qualitative Explanation of Propagation Behavior

The propagation behavior of the slot waveguide may be understood by referring to the transverse equivalent network for the structure, shown in Fig. 4, together with the curves in Figure 3. For complete guiding, the unplated region must be "slow" and the plated regions "fast"; the fields in the plated regions then decay exponentially in the transverse direction. By inspection of the solid line dispersion curves in Fig. 3, we observe that the TM_m mode is faster (smaller value of n_{eff}) than the TM_u mode. This occurs because the metal plating possesses a negative real dielectric constant (with a small imaginary part); its effect is therefore to lower the net value of dielectric constant for the combination of plating plus film and substrate in the plated regions.

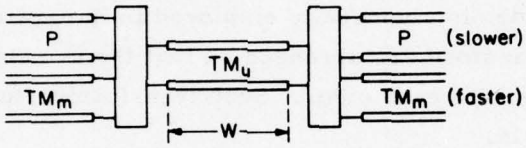


Fig. 4. Transverse equivalent network for the slot waveguide.

Thus, if the P mode could be neglected, the wave would indeed be purely bound, with exponential decay occurring in the plated regions. Under that condition, the transverse equivalent network in Fig. 4 would reduce to a network with one mode inside and only one mode in each of the outside regions. The resulting numerical values for the propagation constant would be almost exactly the same as those obtained in Reference 4. This solution is shown qualitatively in Fig. 3 as the curve with short dashes, labeled "without P." Because of the approximation (only one mode in each region), the curve does not exist below thickness t_2.

In the more accurate representation shown in Fig. 4, the P transmission lines in the outer, metallized regions are included. For thicknesses less than t_2, the TM_m transmission lines are no longer present, but the P lines are, so that propagation can continue down to thickness t_1, the cutoff thickness for the TM_u mode by itself. Inclusion of the P mode thus shows that the slot waveguide can still guide energy even though the film thickness is smaller than t_2. This solution is shown by the line with longer dashes, labeled "with P." It is also seen that for film thicknesses reasonably above t_2 the earlier solution for n_{eff} is quite accurate.

However, the P mode, as seen from Fig. 3, is even slower than the TM_u mode. The P transmission lines in Fig. 4 are therefore above cutoff, with the result that some energy must be leaking transversely as the wave progresses down the slot guide.

The attenuation constant for the slot guide when the P mode is included should therefore be greater than the value computed without the P mode for two reasons: the P mode is intrinsically lossier and introduces new attenuation, and leakage of energy is produced, as shown above.

2. Numerical Values for the Dispersion and Attenuation Behavior

Numerical values for the propagation behavior as a function of certain geometrical parameters were calculated from the transverse equivalent network shown in Figure 4. In this network, the transmission line parameter values are known rigorously from the respective infinitely-wide plated and unplated regions. The boxes, which represent the coupling between the various transmission lines, are treated approximately. Rigorously, the boxes should contain contributions from the continuous spectrum, but in our analysis they are neglected since we expect their influence to be numerically negligible for the very wide slots generally employed in integrated optics. Our boxes actually comprise only transformers arranged so that the two outside transmission lines on each side are in series with each other. Such transformer turns ratios are given by simple overlap integrals.

The structure chosen for analysis consists of film and substrate refractive indices of 1.68 and 1.48, respectively, with a silver plating. Silver has a dielectric constant at a wavelength of 0.633 μ of about -16-j0.5; the accurate refractive index is 0.067-j4.040. The dispersion curve shall be phrased in terms of $n_{eff}(=\lambda/\lambda_g = \beta/k)$ vs. t/λ, where t is the film thickness and $\lambda(=2\pi/k)$ is the free-space wavelength. When the slot is wide relative to wavelength, the dispersion curve is very close to that for the TM_u mode, corresponding to the unplated region itself, because most of the field is then in the unplated region. In order to obtain a curve which deviates from the TM_u values, we show in Fig. 5 the results for a rather narrow slot, for $W/\lambda = 2.0$, where W is the slot width. It is seen that the behavior follows very closely that predicted by the qualitative discussion above. For values of t/λ above the cutoff for the TM_m mode, the curve (shown dashed) neglecting the P mode yields values almost identical to those for the accurate calculation, including the P mode. It should be noted that for most situations the film and substrate refractive indices will be closer to each other than in this calculation, and that the slot widths would be greater; for those situations, the differences between the dashed and the heavy solid curves would really be negligible.

For values of t between the cutoffs of the TM_m and TM_u curves, on the other hand, the dashed and heavy solid curves are entirely different, in that the dashed curve predicts no guided propagation at all in this film thickness range. Thus, the inclusion of the P mode in the analysis shows that guided propagation can occur down to lower film thicknesses than previously believed. The slight kink appearing in the curve labeled

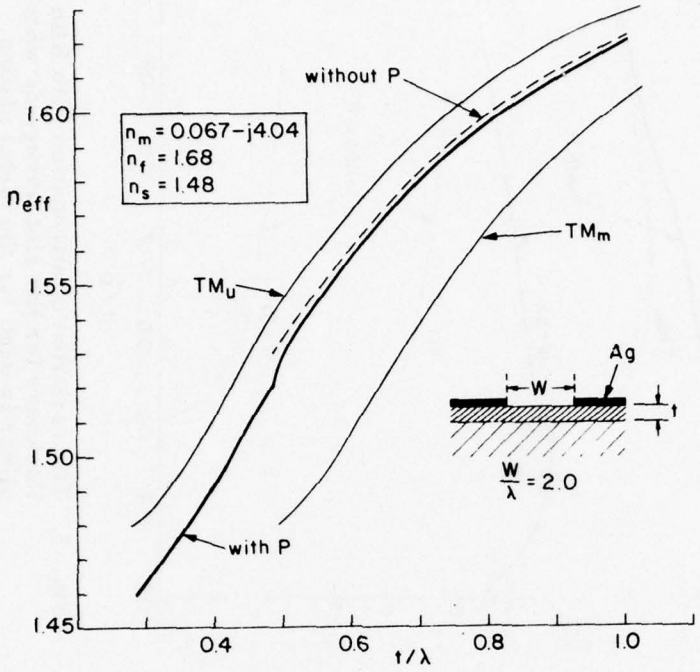


Fig. 5. Dispersion behavior for a slot waveguide employing a silver plating, expressing $n_{eff}(=\lambda/\lambda_g = \beta/k)$ as a function of normalized film thickness, t/λ.

"with P" at the cutoff of the TM_m mode is due, we believe, to the neglect of the continuous spectrum in the analysis.

A curve of attenuation α as a function of slot width W is presented in normalized form in Fig. 6, when the refractive index values are the same as those in Fig. 5 and the film thickness is given by $t/\lambda = 0.60$, in the range for which both outside transmission lines exist in the network of Figure 4. It is seen that the attenuation drops substantially as the slot width is increased, which is expected since then less of the field is present in the metallized regions, and that the inclusion of the P mode increases the attenuation by more than an order of magnitude. For very narrow strips, the attenuation is seen to exhibit a peak and then a decrease, as a result of the changing partitioning of field between the P and the TM_m modes in the outside (plated) regions. We see from Fig. 6 that lower attenuation is obtainable when the slot width is increased, but that nevertheless the inclusion of the P mode in the solution introduces a serious increase in α.

The variation of attenuation with film thickness t is shown in Fig. 7, for two different slot widths. We see first that, consistent with Fig. 6, the attenuation is always

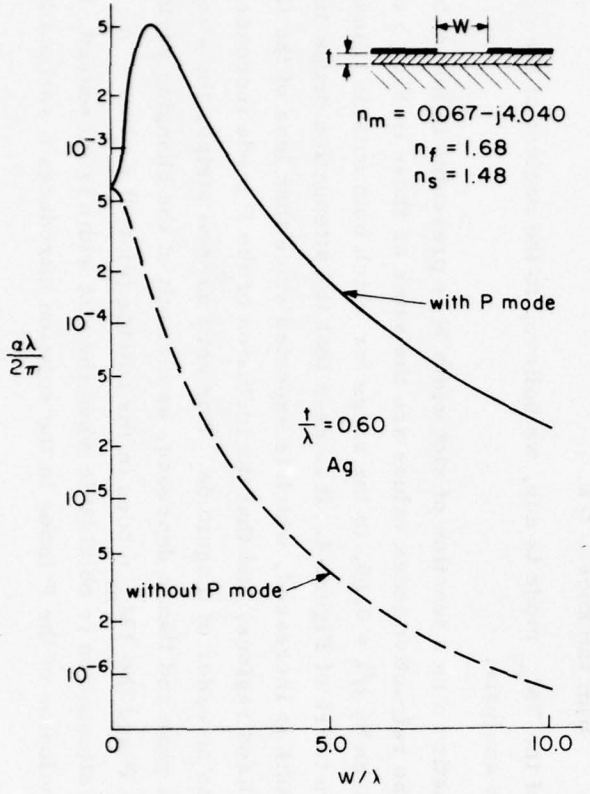


Fig. 6. Attenuation vs. slot width, both normalized, for the slot waveguide when silver is used for the metal plating.

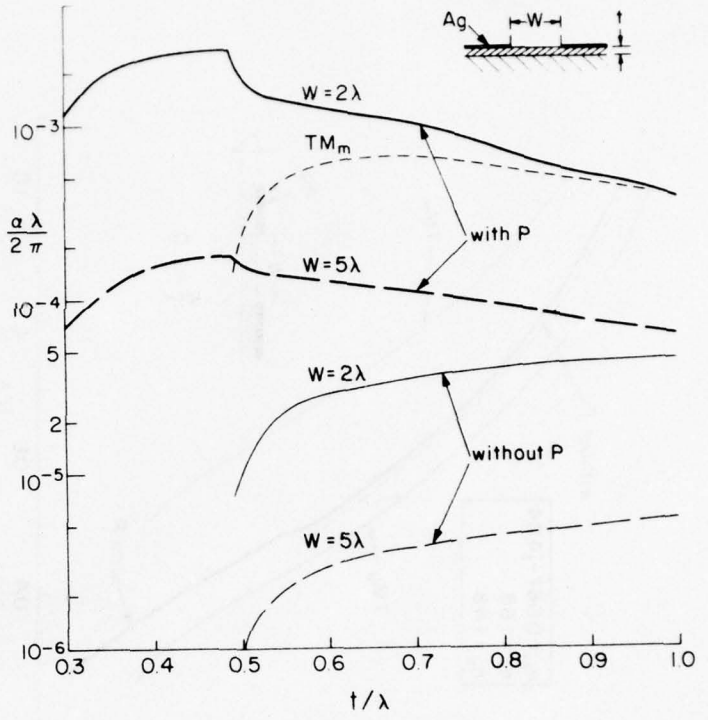


Fig. 7. The variation of attenuation with film thickness for the slot waveguide when silver is used for the metal plating.

higher for smaller values of W. We also observe quite clearly that the solution with P yields substantially higher α values. It is also evident that the solution with P continues down to lower values of t/λ, consistent with Fig. 5, but it is interesting that the α values do not increase in this range. The curve of attenuation for the TM_m mode alone is also shown (small dashes); it is seen that for large W/λ values it is easy to achieve attenuations substantially below these values.

B. The Metal-Strip Waveguide

The structure of this waveguide was shown in Fig. 2; it consists of a metal strip on a thin film on a substrate. The metal is again viewed as an overdense plasma.

The transverse equivalent network for this simple structure, which apparently has not been analyzed previously or recognized as a possible optical waveguide, is shown in Figure 8. For complete guidance, the TM_u mode transmission line must be below cutoff; this requirement is satisfied over all ranges of film thickness because the P mode is always slower than the TM_u mode, as seen from Figure 3. The TM_m mode, on the other hand, is even faster than the TM_u mode, so that its transmission line is also below cutoff. Thus, the only propagating line in Fig. 8 is the P mode line; the wave guided by the metal-strip structure should therefore be dominated by the P mode. Since the P mode is present only for E-vertical polarization, the metal-strip structure has the immediate advantage that only the E-vertical polarization modes can be guided by it.

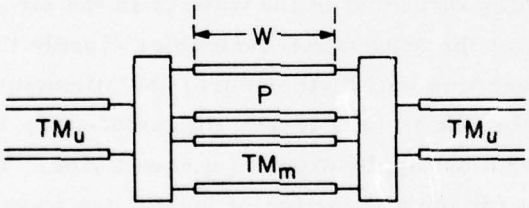


Fig. 8. Transverse equivalent network for the metal-strip waveguide.

The metal strip will also guide for thicknesses smaller than t_1 (and even if the dielectric film is completely absent), although we have not analyzed that range. The metal strip waveguide will therefore guide completely (no leakage, but with high attenuation) over all values of film thickness.

A theoretically-calculated plot of n_{eff} as a function of t/λ for this guiding structure is presented in Fig. 9 for a silver strip and the other parameters chosen previously. As expected, the curve resembles that of the P wave itself, so that the effective refractive index for the wave guided by the strip is substantially higher than that for the

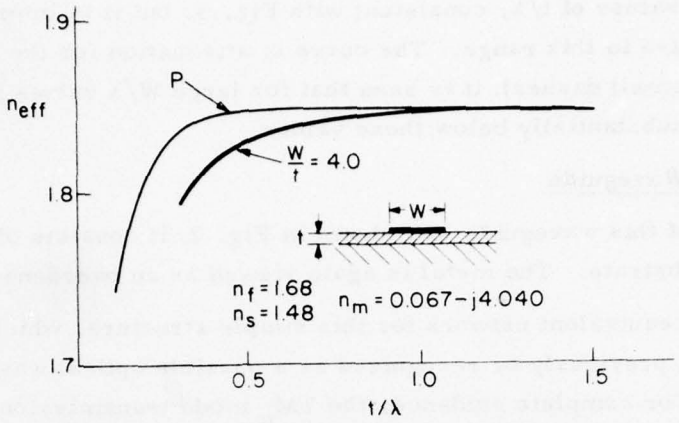


Fig. 9. Dispersion curve (n_{eff} vs. film thickness) for the metal-strip waveguide when silver is used.

unplated region alone. As a result, the wave is more tightly bound to the strip than is the case for other optical linear waveguides. A practical consequence of this tight binding is that it permits the waveguide to undergo much sharper bends before the leakage radiation becomes significant. This problem relating to bends is a serious one for integrated-optical waveguides because the n_{eff} for the waveguide is usually almost the same as that for the neighboring regions, resulting in very weak binding.

Tight binding also implies that most of the wave is in the strip region; as a consequence, we would expect that the attenuation resembles closely that of the P wave alone. Alas, calculations bear this out; furthermore, the attenuation changes little with strip width. Since the P mode is very lossy, the metal-strip waveguide possesses an attenuation value which for most applications is disqualifying. This aspect is very unfortunate, because, except for the attenuation property, the waveguide possesses an unusual set of virtues. It may be possible that the P mode for other materials at other frequencies possesses much lower attenuation values, but we have not investigated this feature.

C. Conclusions

The earlier theoretical analyses[3,4] for the slot waveguide are correct for the H-vertical polarization modes, but incorrect for the E-vertical polarization modes, because a constituent plasmon mode, due to the metal plating, was omitted from the analysis. A transverse equivalent network approach which includes the plasmon mode shows that two major effects occur:

(1) The wave can be guided over a larger range of film thicknesses than previously thought; guiding can occur down to the cutoff thickness of the unplated region.

(2) The wave is not completely guided, as previously thought, but leaks energy transversely. This leakage, plus the higher intrinsic attenuation of the plasmon mode, increases the attenuation of the guided wave by more than an order of magnitude over earlier calculations.

Because of the higher attenuation possessed by the E-vertical polarization modes of the slot waveguide, they should be avoided whenever possible and the H-vertical modes should be used instead. The plasmon mode is not a constituent of the H-vertical modes, so that their attenuation values are significantly lower than those of the E-vertical modes for the same slot widths. Fortunately, the modulator or switch applications projected for the slot waveguide,[7] which involve an rf or dc electric field between the plated regions, automatically requires the use of the H-vertical optical mode (or E^x mode in Marcatili's notation[5]) for optimum interaction.

The metal-strip waveguide is a simple structure which has previously not been analyzed or recognized as a possible optical waveguide. Its only disadvantage, which is a very serious one, is its unusually high attenuation, about that of the plasmon mode by itself. This is very unfortunate because the waveguide possesses several unique advantages, in addition to its very simple structure. Its unique advantages are:

(1) It guides completely (no leakage), and it guides only for the E-vertical polarization, thus eliminating a whole set of higher modes and the usual cross-polarization problems.

(2) Its wave is very tightly bound, thus permitting much sharper bends and overcoming one of the principal problems facing the use of integrated-optical waveguides.

(3) It will guide for all values of film thickness, and even without the presence of a film.

Perhaps sufficiently low attenuation values, permitting this unique waveguide to become practical, may be possible with other materials at other frequencies.

Joint Services Technical Advisory Committee
F44620-74-C-0056

A.A. Oliner and S.T. Peng

REFERENCES

1. A.A. Oliner, "Optical and Acoustical Micro-Electronics: Similarities and Differences," MRI Symp. Proc. XXIII, Optical and Acoustical Micro-Electronics, 1-17 (especially pages 12, 13) (New York: Polytechnic Press, 1974).

2. A.A. Oliner, "Acoustic Surface Waveguides and Comparisons with Optical Waveguides," IEEE Trans. Microwave Theory Tech., 24 (December 1976).

3. J. Hamasaki and K. Nosu, "A Partially Metal-Clad-Dielectric-Slab Waveguide for Integrated Optics," IEEE J. Quantum Electron., 10, 822-825 (October 1974).

4. Y. Yamamoto, T. Kamiya and H. Yanai, "Propagation Characteristics of a Partially Metal-Clad Optical Guide: Metal-Clad Optical Strip Line," Appl. Optics, 14, 322-326 (February 1975).

5. E.A.J. Marcatili, "Dielectric Rectangular Waveguide and Directional Coupler for Integrated Optics," Bell Syst. Tech. J., 48, 2071-2102 (September 1969).

6. I.P. Kaminow, W.L. Mammel and H.P. Weber, "Metal-Clad Waveguides: Analytical and Experimental Study," Appl. Optics, 13, 396-405 (February 1974).

7. J.C. Campbell, F.A. Blum, D.W. Shaw and K.L. Lawley, "GaAs Electro-Optic Directional-Coupler Switch," Appl. Phys. Letters, 27, 202-205 (15 August 1975).

RAY OPTICAL CALCULATION OF EDGE DIFFRACTION IN UNSTABLE RESONATORS

L.B. Felsen and C. Santana

By a fairly recent generalization,[1,2] it has been shown how high frequency ray optical techniques can be adapted to the analysis of scattering by localized discontinuities (small obstacles, edges, etc.) in waveguides or ducts filled with homogeneous or inhomogeneous dielectric media. Basic to the technique is the ray optical formulation of the waveguide Green's function, i.e., the radiation from a source with isotropic radiation pattern. This is then generalized to non-isotropic sources with radiation pattern $f(\theta)$ where θ is the angle measured from the waveguide axis y. (Although three-dimensional problems can be treated by this method, we consider here only the two-dimensional z-independent case.) A localized discontinuity may be characterized by its free-space diffraction pattern $\bar{f}(\theta, \theta_i)$ when the incident field is a uniform plane wave impinging from the direction θ_i. When the discontinuity is placed inside the waveguide and illuminated by an incident waveguide mode, which can locally be decomposed into uniform plane waves with characteristic angles $\theta_j^{\pm}$, the resulting $\bar{f}(\theta, \theta_j^{\pm})$ constitutes an equivalent non-isotropic source whose excitation of modal fields may be calculated from the solution referred to above. In this manner, one derives by ray optical techniques the modal reflection, transmission and coupling coefficients (i.e., the scattering matrix elements) for a discontinuity inside the waveguide. The lowest-order, single diffraction solution so obtained may be refined by accounting for multiple diffraction effects due to interaction between the singly diffracted fields and the waveguide boundaries. For details of the method, the reader is referred to previous work.[2-6]

The ray optical technique has already been applied to the study of discontinuities in waveguides of various types.[6,7] In the present report, it is shown how it can be applied to the important problem of unstable open optical resonators. Because of their good mode selectivity and large mode volume, such structures appear to be most promising for use with laser sources of high and even moderate gain.[8] By recent studies performed independently in the United States[9,10] and the Soviet Union,[11] it has been shown that the unstable resonator can be regarded as a waveguide whose boundaries are the convex resonator mirrors and whose axis is transverse to the resonator axis (Figure 1). Resonance in this open waveguide is established by self-consistent reflection of a propagating waveguide mode between the edge discontinuities formed by the rims of the mirrors. Although the waveguide is very strongly overmoded, it has been shown that near the cutoff condition, which is of interest for the resonator problem, mode coupling due to the mirror edges is confined essentially to adjacent modes. Thus, a very simple model involving selective coupling between two waveguide modes has been developed and has been found capable of explaining the intricate eigenmode loss behavior

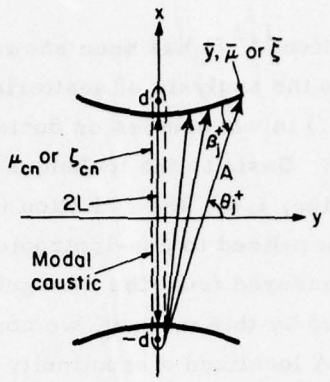


Fig. 1. Unstable resonator with hyperbolic mirrors. The resonator axis lies along x and the waveguide axis along y. A typical modal caustic, an ellipse with foci at $x = \pm d$, is shown, together with one congruence of upgoing modal rays. Ray A strikes the upper edge, and θ_j^+ is the propagation angle of the corresponding local plane wave field in the j-th mode. A similar congruence of downgoing modal rays has been omitted, as has the corresponding picture for the left half of the resonator. For very slender caustics, all modal rays appear to originate at the foci. The edges of the mirrors are located at $\pm \bar{y}$, $\bar{\mu}$ or $\bar{\zeta}$ in the various coordinate systems defined in the text; the analogous designation for the modal caustic is μ_{cj} or ζ_{cj}.

determined by numerical solution of the resonator integral equation.[9,10] While the role of mode coupling has been alluded to in the Russian work,[11] it has not been incorporated into their analysis. The Soviet calculations are based on a single mode analysis, which is adequate only near eigenmode loss minima and does not provide the peculiar interconnections between successive loss minima found in the numerical results.

A further attribute to our analysis[9,10] is the avoidance of the resonator integral equation, which forms the basis of the Soviet approach to the waveguide problem as originally formulated by Weinstein[12] and followed thereafter by others in the Soviet Union.[11,13] By avoiding the integral equation, it is possible to decompose the unstable resonator problem into conventional microwave network constituents involving propagation (waveguide) and discontinuity regions. By this separation, one may extend the analysis also to resonator configurations which are filled with inhomogeneous and(or) active materials, and to mirror shapes which depart from the conventional circular contours.

While the ray optical principle of localization is consonant with the microwave network approach, the reflection and coupling coefficients due to the mirror edges were

previously[9,10] not calculated by the ray optical method described earlier. Instead, these coefficients were taken from Weinstein[12] by modeling the region near the edges locally as an open-ended parallel plane waveguide. Since the ray optical method synthesizes the reflection and coupling coefficients by direct edge scattering, it is of interest to examine whether the two procedures yield the same result. This is especially important for resonators with moderately large Fresnel numbers where the local parallel plane approximation near the edges is more difficult to justify since the slanting of the convex waveguide boundaries is then not negligible. (The Fresnel number is defined as $N = k\bar{y}^2/4\pi L$; k is the wavenumber and $\bar{y}$ and L are given in Figure 1.) It will be shown that the single-diffraction ray-optically evaluated reflection and coupling coefficients are identical with those obtained by Weinstein from the rigorous solution of the open-ended parallel plane waveguide problem when the characteristic angle of the incident mode is not almost 90° with respect to the waveguide walls; this is the range of interest for the moderately large Fresnel number regime. This confirmation then suggests that the single diffraction ray optical model may be used with confidence for large and very large Fresnel numbers where the local parallel plane approximation is clearly in doubt. The reflection and coupling coefficients derived here may then be regarded as more reliable than any of those available heretofore. Although the ray optical edge diffraction mechanism has been proposed as an explanation of the numerically observed eigenmode loss behavior,[14,16] this fact has not been incorporated into a systematic modal theory. Such an incorporation is performed here.

A. The Waveguide Green's Function

We seek a solution of the equation

$$(\nabla^2 + k^2)\, G(\underset{\sim}{\rho}, \underset{\sim}{\rho}') = -\delta(\underset{\sim}{\rho} - \underset{\sim}{\rho}') \quad , \quad \underset{\sim}{\rho} = (\mu, \eta) \tag{1}$$

subject to the boundary conditions

$$\frac{\partial G}{\partial \mu} = 0 \quad \text{at} \quad \mu = 0 \tag{1a}$$

$$G = 0 \quad \text{at} \quad \eta = \pm \eta_1 \quad , \tag{1b}$$

and a radiation condition at $\mu \to \infty$. A time dependence $\exp(-i\omega t)$ is suppressed. Here, μ and η are constant coordinate surfaces in an elliptic coordinate system (Figure 2). The boundary condition in Eq. (1b) identifies G as the single component electric field $E \equiv E_z$ and the source as a suitably normalized line of z-directed electric currents. Since we shall be interested only in field solutions which are even with respect to the $y = 0$ plane, the boundary condition in Eq. (1a) has been imposed and effectively bisects the waveguide.

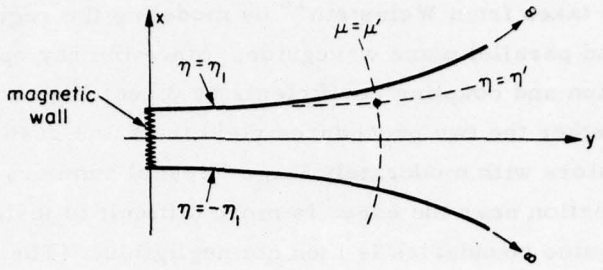


Fig. 2. Bisected waveguide configuration.

In terms of waves propagating along the y (or μ) directions, the Green's function can be represented as:

$$G(\underline{\rho}, \underline{\rho}') = \sum_n \frac{\Phi_n(\eta)\,\Phi_n(\eta')}{N_n^2}\, g_n(\mu, \mu') \tag{2}$$

where

$$N_n^2 = \int_{-\eta_1}^{\eta_1} \Phi_n^2(\eta)\, d\eta \tag{2a}$$

is the squared normalization constant for the eigenfunctions $\Phi_n(\eta)$. The latter satisfy the source-free one-dimensional equation

$$\left[\frac{d^2}{d\eta^2} + h^2(b_n^2 - \sin^2\eta)\right] \Phi_n(\eta) = 0 \quad , \quad h = kd \tag{3}$$

with b_n representing the modal eigenvalue and

$$\Phi_n(\pm\eta_1) = 0 \tag{3a}$$

The one-dimensional Green's function $g_n(\mu, \mu')$ satisfies the source-excited equation

$$\left[\frac{d^2}{d\mu^2} + h^2(\cosh^2\mu - b_n^2)\right] g_n(\mu, \mu') = -\delta(\mu - \mu') \tag{4}$$

with

$$\frac{dg_n}{d\mu} = 0 \text{ at } \mu = 0, \text{ radiation condition at } \mu \to \infty \tag{4a}$$

The eigenfunctions $\Phi_n(\eta)$ were determined previously.[9] Assuming that the turning

points $\eta_{cn} = \pm \sin^{-1} b_n$ lie outside the domain $-\eta_1 < \eta < \eta_1$ (this is the case for the modes of interest), and since h is large, one may employ the WKB approximations for $\Phi_n(\eta)$:

$$\Phi_n(\eta) \sim \frac{\sin\left[h \int_{\eta}^{\eta_1} \psi_n(\tau)\, d\tau\right]}{\left[\psi_n(\eta)\right]^{1/2}}, \quad \psi_n(\eta) = (b_n^2 - \sin^2 \eta)^{1/2} \tag{5}$$

with b_n defined by Eq. (3a) via

$$h \int_{-\eta_1}^{\eta_1} \psi_n(\tau)\, d\tau = n\pi \ , \quad n = \text{integer} \tag{5a}$$

or approximately as,[9]

$$b_n^2 \approx 2\, \frac{n\pi - 2kL}{k \ln M} + 1 \tag{5b}$$

where

$$M = 1 + 2\gamma + 2\sqrt{a} \ , \quad a = \gamma + \gamma^2 \ , \quad \gamma = L/r \tag{5c}$$

M is the linear magnification and γ is the ratio of the resonator half length L measured along the x axis to the radius of curvature r of the mirrors on this axis. When the sine function Eq. (5) is expressed as the sum of two exponentials and the result then substituted into Eq. (2a), one observes that two of the resulting integrals contribute negligibly because of rapidly fluctuating integrands. Thus,

$$N_n^2 \sim \frac{1}{2} \int_{-\eta_1}^{\eta_1} \frac{d\tau}{\psi_n(\tau)} \approx \frac{\ln M}{2} - (b_n^2 - 1)\left(\frac{\ln M}{8} + \frac{a^{1/2}}{4}\right) \tag{6}$$

The evaluation of the integral in Eq. (6) is based on the approximation $b_n \approx 1$, which applies to the modes of interest (see Reference 9).

The modal Green's function $g_n(\mu, \mu')$ can be constructed in terms of solutions which satisfy the source-free Eq. (4) and the boundary conditions at $\mu = 0$ and $\mu \to \infty$, respectively.[17] Introducing

$$\zeta = (2h)^{1/2} \mu \, , \quad \hat{g} = (2h)^{1/2} g, \quad p_n = \frac{h}{2}\, \mu_{cn}^2 = \frac{h}{2}(b_n^2 - 1) \tag{7}$$

where μ_{cn} locates the turning point (modal caustic) of the approximated differential equation (4), one may write

$$\left(\frac{d^2}{d\zeta^2} + \frac{\zeta^2}{4} - p_n\right) \hat{g}_n(\zeta, \zeta') = \delta(\zeta - \zeta') \tag{8}$$

with boundary conditions corresponding to Equation (4a). The solution for $\hat{g}_n$ is:

$$\hat{g}_n = e^{\pi p_n/2}(2^{ip_n})\frac{\Gamma(\frac{1}{4}+i\frac{p_n}{2})}{\Gamma(\frac{1}{4}-i\frac{p_n}{2})}\left\{D_\nu(z_<)+i(2^{-ip_n})\frac{\Gamma(\frac{1}{4}-i\frac{p_n}{2})}{\Gamma(\frac{1}{4}+i\frac{p_n}{2})}D_{-\nu-1}(iz_<)\right\}D_\nu(z_>) \tag{9}$$

where

$$z = \zeta e^{-i\pi/4}, \quad \nu = -1/2 - ip_n \tag{10}$$

For observation points $\zeta >> |\zeta_{cn}|$, $\zeta_{cn} = 2|p_n|^{1/2}$, i.e., far enough from the modal caustic at ζ_{cn}, one may employ the WKB approximations for the parabolic cylinder functions to obtain

$$\hat{g}_n \sim \frac{\exp[i\int_{\zeta_<}^{\zeta_>}\varphi_n(\xi)\,d\xi]}{-2i[\varphi_n(\zeta)\,\varphi_n(\zeta')]^{1/2}} + B_n\frac{\exp[i\int_{\zeta_<}^{\zeta_>}\varphi_n(\xi)\,d\xi + 2i\int_{\zeta_{cn}}^{\zeta_<}\varphi_n(\xi)\,d\xi]}{-2i[\varphi_n(\zeta)\,\varphi_n(\zeta')]^{1/2}} \tag{11}$$

where

$$\varphi_n(\zeta) = (\frac{\zeta^2}{4} - p_n)^{1/2} \tag{12}$$

and

$$B_n = \frac{\exp(-i\pi/4)}{2^{-ip_n}}\frac{\Gamma(\frac{1}{4}+i\frac{p_n}{2})}{\Gamma(\frac{1}{4}-i\frac{p_n}{2})}\exp[ip_n(1-\ln p_n)] \tag{13}$$

The first term in Eq. (11) represents the direct contribution from the source point at ζ' to the observation point at ζ, while the second term describes a wave that has traveled from the source point to the modal caustic and thence to the observation point. The reflection coefficient due to the caustic, as seen from the observation point, is given by the ratio of the reflected and incident waves as:

$$\tilde{R}_n(\zeta) \sim B_n \exp[2i\int_{\zeta_{cn}}^{\zeta}\varphi_n(\xi)\,d\xi] \tag{14}$$

which has been given previously.[10]

When the results in Eqs. (5) and (11) are substituted into Eq. (2), one obtains the WKB approximated Green's function for $\zeta < \zeta'$:

$$G \sim \sum_n \frac{\sin\left[h\int_{\eta}^{\eta_1}\psi_n(\tau)\,d\tau\right]\sin\left[h\int_{\eta'}^{\eta_1}\psi_n(\tau)\,d\tau\right]}{N_n^2[\psi_n(\eta)\psi_n(\eta')]^{1/2}} \cdot \frac{\exp\left(-i\int_{\zeta}^{\zeta}\varphi_n(\xi)\,d\xi\right) + B_n\exp\left[-i\int_{\zeta'}^{\zeta}\varphi_n(\xi)\,d\xi + 2i\int_{\zeta_{cn}}^{\zeta}\varphi_n(\xi)\,d\xi\right]}{-2i(2h)^{1/2}[\varphi_n(\zeta)\varphi_n(\zeta')]^{1/2}} \tag{15}$$

B. Response to a Directive Source

The Green's function in Eq. (15) represents the field excited by a line source of strength

$$G_0(|\underset{\sim}{\rho} - \underset{\sim}{\rho}'|) = \frac{i}{4}H_0^{(1)}(k|\underset{\sim}{\rho} - \underset{\sim}{\rho}'|) \sim \frac{i}{4}\left(\frac{2}{\pi k|\underset{\sim}{\rho} - \underset{\sim}{\rho}'|}\right)^{1/2}\exp[ik|\underset{\sim}{\rho} - \underset{\sim}{\rho}'| - i\pi/4] \tag{16}$$

When the source has a radiation pattern $f(\theta)$ so that its far field is

$$\bar{G}_0 \sim G_0\, f(\theta) \tag{17}$$

the response in the waveguide can be calculated by decomposing the source-point dependent eigenfunction $\Phi_n(\eta')$ into its local plane wave constituents and multiplying these by $f(\pi + \theta_n'^{\pm})$, where $\theta_n'^{\pm}$ are the propagation angles of the local plane waves in the n-th mode at the source point. Thus, the field $\bar{G}$ at $\zeta < \zeta'$ produced when the directive source in Eq. (17) is placed inside the waveguide is given by Eq. (15) provided that one makes the replacement

$$\sin\left[h\int_{\eta}^{\eta_1}\psi_n(\tau)\,d\tau\right] \to (2i)^{-1}\left\{f(\pi+\theta_n'^{-})\exp\left[ih\int_{\eta}^{\eta_1}\psi_n(\tau)\,d\tau\right] - f(\pi+\theta_n'^{+})\exp\left[-ih\int_{\eta}^{\eta_1}\psi_n(\tau)\,d\tau\right]\right\} \tag{18}$$

If the source is located on the lower wall or the upper wall, one puts $f(\pi + \theta_n'^{+}) = 0$ and $f(\pi + \theta_n'^{-}) = 0$, respectively[2]; the pattern functions now represent the respective far zone fields of the source in the presence of the boundary whereon it is situated.

C. Reflection Coefficient for the Open-Ended Waveguide

1. General Formulation for Large Fresnel Numbers

When the waveguide in Fig. 2 is truncated at $\zeta = \bar{\zeta}$, reflection occurs from the open end. By the ray optical method, the modal reflection and coupling coefficients

are calculated from the single edge diffraction patterns. This requires first a determination of the local plane wave fields that illuminate the edges. Then, one calculates the diffraction field due to the upper edge, observed at an angle β with respect to the tangent plane at the edge. This provides the strength and pattern of the equivalent line source in Eq. (17), which is then inserted into Equations (18) and (15). Omitting the term multiplied by B_n in Eq. (15), which represents fields non-relevant for edge reflection (they arrive after a round trip to the center of the resonator), one obtains for the total reflected field E_{rj} due to an incident mode j:

$$E_{rj} \sim \sum_n \Gamma_{jn} \left\{ N_n [\psi_n(\eta)\, \varphi_n(\zeta)]^{1/2} \right\}^{-1} \sin\left[h \int_{\eta}^{\eta_1} \psi_n(\tau)\, d\tau\right] \exp\left[-i \int_{\zeta_{cn}}^{\zeta} \varphi_n(\xi)\, d\xi\right] \quad (19)$$

where

$$\Gamma_{jn} = \frac{[1 + (-1)^{n+j}]\, V(\beta_n, \beta_j)}{8i(2h)^{1/2} [\psi_n(\eta_1)\, \psi_j(\eta_1)\, \varphi_n(\bar{\zeta})\, \varphi_j(\bar{\zeta})]^{1/2} N_n N_j} \cdot \exp\left[i \int_{\zeta_{cn}}^{\bar{\zeta}} \varphi_n(\xi)\, d\xi + i \int_{\zeta_{cj}}^{\bar{\zeta}} \varphi_j(\xi)\, d\xi\right] \quad (20)$$

Here, Γ_{jn} is the coupling coefficient, due to diffraction at both edges, from the incident mode j to mode n, and $V(\beta_n, \beta_j)$ is the edge diffraction coefficient for a perfectly reflecting half plane illuminated by a unit strength plane wave incident at the angle β_j and observed at the angle β_n:

$$V(\beta_n, \beta_j) = -\sec \frac{\beta_n - \beta_j}{2} + \sec \frac{\beta_n + \beta_j}{2}, \quad \beta_n \pm \beta_j \neq \pi \quad (21)$$

β_n and β_j denote the characteristic propagation angles of the local plane waves that comprise modes n and j.

2. Simplification for Moderate Fresnel Numbers

To check whether the general coupling coefficient in Eq. (20) reduces to the local parallel plane approximation for small enough Fresnel numbers and in particular for weakly curved mirrors (small γ in Eq. (5c)), the limiting form of Eq. (20) in that parameter range is now examined. As noted earlier, the modes of interest are those with caustics near the center of the resonator. Thus, the edges may be taken to lie far enough from the caustics to justify $\mu_{cn,j} \ll \bar{\mu}$. Moreover, $\bar{\mu}$ itself is small. When these approximations are introduced, one finds:

$$V(\beta_n, \beta_j) \approx (\operatorname{sgn} \beta_n)(\operatorname{sgn} \beta_j) \frac{4\bar{\mu} \cos \eta_1}{(2\bar{\mu}^2 + 1 - b_n^2) + (2\bar{\mu}^2 + 1 - b_j^2)} \tag{22}$$

$$\ln M \approx 2\gamma^{1/2}, \quad a \approx \gamma, \quad N_{n,j}^2 \approx a^{1/2} \tag{23}$$

$$\bar{\mu}\left(1 - \frac{\mu_{cn}^2}{2\bar{\mu}^2}\right) \approx (\bar{\mu}^2 - \mu_{cn}^2)^{1/2} = (2/h)^{1/2} \varphi_n(\bar{\zeta}) \tag{24}$$

Thus, Eq. (20) reduces to, when the phase reference is at the edges,

$$\Gamma_{jn} \approx -i[(s_n + s_j)(s_n s_j)^{1/2}]^{-1}, \quad s_n = 2a^{1/4} \varphi_n(\bar{\zeta}) \tag{25}$$

and

$$\Gamma_{jj} \approx -\frac{i}{2s_j^2} \tag{26}$$

These expressions agree with the parallel plane formulas of Weinstein[12] for the case where β_n and β_j differ sufficiently from $\pi/2$ to ensure that one edge does not lie in the reflected ray transition region of the other. In that event, single-edge diffraction is adequate to describe reflection from the open-ended parallel plane structure. As noted previously,[9] s_n represents the modal propagation coefficient in the equivalent parallel plane waveguide whose height equals that of the hyperbolic waveguide at $\bar{\zeta}$.

When γ is reduced further so that one edge does lie in the reflected ray transition region of the other edge, interaction between the edges in the local parallel plane model cannot be ignored. The single-edge diffraction function must now be replaced by a more accurate function derived from the rigorous solution of the semi-infinite parallel plane configuration. The result as given by Weinstein,[20] modified to account for the different normalization of the incident field used here, is as follows:

$$\Gamma_{jn} = -i[(s_n + s_j)(s_n s_j)^{1/2}]^{-1} \exp[U(s_n, \delta) + U(s_j, \delta)] \tag{27}$$

where δ is defined by any of the equalities

$$\frac{2\pi(\hat{j} + \delta)}{\ln M} = p_n = \frac{h}{2}(b_n^2 - 1) = \frac{n\pi - 2kL}{\ln M}, \quad \hat{j} = 0, \pm 1, \pm 2, \cdots \tag{28}$$

The diffraction function $U(s, \delta)$ is discussed in detail in Ref. 20 and has the following asymptotic behavior:

$$U(s, \delta) \sim 0, \quad s \text{ large} \tag{29}$$

$$U(s, \delta) \sim \frac{1}{2}\left(\ln 2 + \frac{i\pi}{2}\right) + \ln(2s) - \frac{1 - i}{2}\bar{\beta}s + \cdots, \quad s \text{ small} \tag{30}$$

where $\bar{\beta} = 0.824$. Thus, Eq. (27) reduces to Eq. (25) when $s_{n,j}$ are sufficiently large (note that this condition can be met although γ is small). For small $s_{n,j}$, one has

$$\Gamma_{jn} = -2(s_n + s_j)^{-1}(s_n s_j)^{1/2} \exp[-(1-i)\bar{\beta}(s_n + s_j)^{1/2}] \tag{31}$$

which (noting that $s_n \approx s_j$ here) was used in the analysis by Chen and Felsen.[9] With formulas (20) and (27), noting the overlap region in Eq. (25), one may cover the entire range of parameters from small to large Fresnel numbers for the hyperbolic mirror resontaor.

This completes the objective stated at the beginning of this report.

Joint Services Technical Advisory Committee
F44620-74-C-0056 L.B. Felsen and C. Santana

National Science Foundation
ENG75-22625

REFERENCES

1. H.Y. Yee, L.B. Felsen and J.B. Keller, "Ray Theory of Reflection from the Open End of a Waveguide," SIAM J. Appl. Math., 16, 268-300 (1968).
2. D.V. Batorsky and L.B. Felsen, "Ray-optical Calculation of Modes Excited by Sources and Scatterers in a Weakly Inhomogeneous Duct," Radio Sci., 6, 911-923 (1971).
3. J.J. Bowman, "Comparison of Ray Theory with Exact Theory for Scattering by Open Waveguides," SIAM J. Appl. Math., 18, 818-829 (1970).
4. J. Boersma, "Ray-optical Analysis of Reflection in an Open-Ended Parallel-Plane Waveguide: II - TE Case," Proc. IEEE, 62, 1475-1481 (1974).
5. S.W. Lee, "Ray Theory of Diffraction by Open-Ended Waveguides: I. Field in Waveguides," J. Math. Phys., 11, 2830-2850 (1970).
6. L.W. Chen and L.B. Felsen, "Ray Method for Scattering by Discontinuities in a Multiwave Waveguide," SIAM J. Appl. Math., 27, 138-158 (1974).
7. D.V. Batorsky and L.B. Felsen, "Scattering by Obstacles in an Inhomogeneously Filled Waveguide," Radio Sci., 8, 547-557 (1973).
8. A.E. Siegman, "Unstable Optical Resonators for Laser Applications," Proc. IEEE, 53, 277-286 (1965).
9. L.W. Chen and L.B. Felsen, "Coupled-Mode Theory of Unstable Resonators," IEEE J. Quantum Electron., QE-9, 1102-1113 (1973).
10. C. Santana and L.B. Felsen, "Two-Dimensional and Three-Dimensional Unstable Resonator Losses by a Waveguide Analysis," Appl. Opt., 15, No. 6, 1470-1478 (1976).
11. G.N. Vinokurov, V.V. Lyubimov and I.B. Orlova, "Investigation of the Selective Properties of Open Unstable Cavities," Opt. Spectrosc. (USSR), 34, 427-432 (1973).
12. L.A. Weinstein, "Open Resonators and Open Waveguides," (Boulder, Colorado: Golem Press, 1969).
13. V.V. Lyubimov and I.B. Orlova, "Approximate Calculation of Oscillations in Resonators with Concave Mirrors," Opt. Spectrosc. (USSR), 29, 310-313 (1970).

14. Yu.A. Anan'ev and V.E. Sherstobitov, "Influence of the Edge Effects on the Properties of Unstable Resonators," Sov. J. Quantum Electron., 1, 263-267 (1971).

15. V.E. Sherstobitov and G.N. Vinokurov, "Properties of Unstable Resonators with Large Equivalent Fresnel Numbers," Sov. J. Quantum Electron., 2, 224-229 (1972).

16. G.L. McAllister, W.H. Steier and W.B. Lacina, "Improved Mode Properties of Unstable Resonators with Tapered Reflectivity Mirrors and Shaped Apertures," IEEE J. Quantum Electron., QE-10, 346-355 (1974).

17. L.B. Felsen and N. Marcuvitz, "Radiation and Scattering of Waves," (Englewood Cliffs, N.J.: Prentice Hall, Inc., 1973), Sec. 3.3.

18. H. Buchholz, "The Confluent Hypergeometric Function," (New York: Springer-Verlag, Inc., 1969).

19. Ibid, Ref. 17, Sec. 6.5.

20. L.A. Weinstein, "The Theory of Diffraction and the Factorization Method," Appendix B and Prob. 1.6, (Boulder, Colorado: Golem Press, 1969).

ANALYSIS AND EXPERIMENTAL EVALUATION OF LOSSY GRATING COUPLERS

S. Austin, S.T. Peng, F.T. Stone and T. Tamir

Much theoretical work has been published on optical grating couplers, but the experimental verification of these results has not kept pace with this theoretical progress. We therefore discuss here the outcome of a joint theoretical and experimental effort wherein the performance of lossy grating couplers has been studied by combining analytical methods with laboratory measurements. In particular, we have evaluated the performance of grating couplers made holographically in Shipley AZ 1350B photoresist, which may incur substantial losses in coupling efficiency if they are not properly designed. However, we have found that careful design and fabrication can minimize scattering and absorption losses so as to realize coupling efficiencies that are close to those predicted by theoretical considerations.

To produce the desired periodic structure for the grating-coupler device, we use an appropriate initial photoresist thickness obtained from calibration curves of resist thickness vs. spinner speed and resist concentration.[1] Then the experimentally determined etch depth vs. exposure characteristic can be used together with the linear dependence of etch depth on development time[2] to construct the coupler. It is helpful to computer-generate families of grating profiles as a function of the various parameters. The profile nearest the desired one can then be selected, thereby fixing the experimental variables. The above procedure has been verified by comparing measured and predicted values of grating diffraction efficiencies.

The finished grating structure generally exhibits a five-layer configuration as shown in Fig. 1, i.e., it consists of a substrate, a waveguide (film) of thickness t_f, a

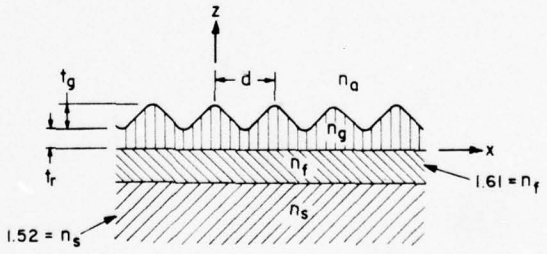


Fig. 1. Five-layer grating structure.

residual photoresist layer of thickness t_r, a grating having a height t_g and an upper air region. For a variety of reasons, it is important to control the grating profile and, in particular, to achieve $t_r = 0$. Thus, to obtain the highest available coupling efficiency for incident beams having Gaussian profiles, it is well known[3] that the grating and

beam parameters must satisfy $\alpha w_o = 0.68$, where α is the imaginary (leakage) part of the propagation constant and w_o is the beam half-width measured along the waveguide. As the value of α depends on all the geometrical and physical parameters of the grating configuration in Fig. 1, their judicious choice is important if a practical beam-width satisfying $w_o = 0.68/\alpha$ is desired. In addition, the absorption losses in the photoresist are expected to degrade the performance of the grating coupler. As the photoresist is required essentially only to provide a periodic profile, we expect that the presence of a residual layer of thickness t_r could have a deleterious effect on performance.

To both determine a suitable value for α and to verify the conjecture that $t_r = 0$ is an optimal condition, we have examined the propagation characteristics of the five-layered structure by using a perturbation procedure,[4] the accuracy of which was verified by a rigorous approach.[5] The calculations thus obtained have guided us in designing the grating structures that were subsequently used in the experimental studies. Furthermore, these calculations have confirmed the expectation that the presence of any residual resist layer does indeed degrade coupler performance. We also found that this degradation is especially strong when the waveguide refractive index n_f is smaller than the real part n_g' of the refractive index n_g of the photoresist material, in agreement with the observations of Dalgoutte and Wilkinson.[6]

In fact, because of the presence of non-negligible losses, our theoretical considerations show that the significant decay parameter is no longer the leakage constant α. Instead, we must use an effective (total) parameter $\alpha_t = \alpha + \alpha_{abs}$ where α_{abs} represents decay due to absorption losses. Thus when α_{abs} is comparable in magnitude to α, the optimum value of w_o for maximum coupling efficiency may be appreciably different from the value determined by the relationship $\alpha w_o = 0.68$.

To confirm these considerations, we have fabricated a number of grating couplers by following the procedures described above and have subsequently compared the leakage and coupling performance to that predicted theoretically. Using the gratings as <u>output</u> couplers, α_t can be obtained from a scan of the output beam profile. The optimum beamwidth w_o can alternatively be obtained by using the grating as an <u>input</u> coupler and examining the coupling efficiency as a function of beamwidth. The formula $\alpha_t = 0.68/w_o$ yields[7] a second value for α_t. A close correlation has been found between these two experimental values of α_t and that predicted from theory.

Gratings were intentionally fabricated with various residual resist layers to study the effect of loss. Table I gives the results for several samples. From this table, we can draw the following conclusions:

TABLE I

t_g (Å)	t_r (Å)	Power Split air(η_a) sub(η_s)	Peak Efficiency (Th)	Peak Efficiency (Exp)	w_o (μm)	α $(\mu m)^{-1}$ (Theory)	$\alpha_t(\mu m)^{-1}$ (Th)	$\alpha_t(\mu m)^{-1}$ (Experiment) (eff)	$\alpha_t(\mu m)^{-1}$ (Experiment) (scan)
800	3000	η_a=.47 η_s=.53	37.6%	31 %	79	.0035	.0076	.0086	.0081
2100	1700	η_a=.50 η_s=.50	40 %	33.5%	102	.0059	.0075	.0067	.0082
3100	700	η_a=.47 η_s=.53	37.6%	37 %	103	.0072	.0077	.0066	*

*not suitable as output coupler

Note: $d = 3160$ Å, $\lambda = 6328$ A, $n''_g = .001$

(1) The measured values of α_t agree with those calculated if the imaginary part of the resist index is taken to be $n''_g = 0.001$

(2) As the thickness of the residual resist layer becomes smaller, the effect of loss becomes negligible, i.e., loss in the grating itself is apparently not a factor

(3) The value of α_t can be obtained by adding the absorption loss factor α_{abs} obtained from an analysis of the uniform five-layer waveguide to the leaky wave α obtained from a lossless analysis of the grating structure.

Input coupling efficiencies and the power split between the two beams at an output coupler also agreed quite well with predicted results. Power split factors of 50% - 50% were usually expected and measured. Using this value of power split, an input coupling efficiency of 0.5 x 80% = 40% is predicted. Values around 35% were obtained using the dip method.

In conclusion, we have demonstrated a procedure for fabricating gratings in photoresist to pre-determined specifications. The theoretically expected and experimentally measured properties agree well if the effect of photoresist loss is included. The effect of loss is due almost entirely to the residual resist layer and, for resist refractive index larger than film index, the effect becomes appreciable if a residual layer is present even though its thickness t_r is equal to only a small fraction of the film thickness t_f.

Joint Services Technical Advisory Committee
F44620-74-C-0056 S. Austin, S. Peng, F. Stone and T. Tamir

Office of Naval Research
N00014-75-C-0421

National Science Foundation
ENG74-23908

REFERENCES

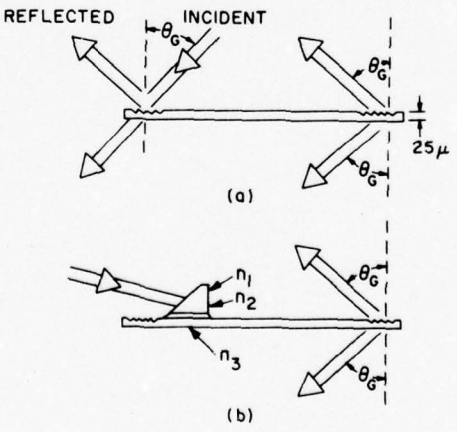
1. S. Austin and F.T. Stone, "Fabrication of Thin Periodic Structures in Photoresist: A Model," Applied Optics, 15, 1071-1074 (April 1976).
2. R.A. Bartolini, Applied Optics, 13, 129 (1974).
3. T. Tamir, "Integrated Optics," 103-107 and 110-112 (Springer-Verlag, 1975).
4. K. Handa, S.T. Peng and T. Tamir, Applied Physics, 5, 325 (1975) and 7, 35 (1975).
5. S.T. Peng, T. Tamir and H.L. Bertoni, IEEE Trans., MTT-23, 123 (1975).
6. D.G. Dalgoutte and C.D. Wilkinson, "An Experimental and Theoretical Study of Thin Grating Couplers for Integrated Optics," to be published in Applied Optics.
7. R. Ulrich, J. Opt. Soc. Am., 63, 1419 (1973).

EFFECTS OF IMPERFECTIONS ON THE EFFICIENCY OF OPTICAL COUPLERS

D.W. Fradlin, P.K. Cheo, S.T. Peng and T. Tamir

The efficient operation of thin-film devices requires that a substantial fraction of the input laser power be coupled through the waveguide and into a useful output beam. When high-quality waveguide structures are used, efficient coupling can be achieved with either a prism or a grating coupler. Work with free-standing (slab) waveguides used in an early model of a thin-film modulator for CO_2 laser radiation[1] has shown both that practical waveguide structures can have serious structural imperfections in the coupling regions and that device constraints may preclude the use of optimized waveguide and coupler designs.

A theoretical study of the effects of structural imperfections on the performance of grating and prism couplers has been conducted. This study was motivated by experiments with symmetric GaAs waveguides that were thinned by chemo-mechanical and ion-mill techniques. It is shown that structural imperfections such as waveguide thickness variations in the coupling region and grating groove depth variations can significantly reduce the coupling efficiency. The partition of energy between two beams above and below the gratings, which represents a serious loss mechanism for the grating coupler, is analyzed for the symmetric GaAs waveguide. The manner in which thickness variations in the coupling region alter this partition is discussed. Data which tend to confirm the predictions of the theory are summarized, and the results of a direct experimental comparison between grating and prism couplers are given (see Figure 1). It is shown that, for the free-standing modulator waveguides the prism



(a) grating couplers, (b) prism input coupler

Fig. 1. Free-standing (slab) waveguide.

coupler is superior because it leads to both enhanced coupling efficiency and enhanced output beam quality. This work establishes fabrication tolerances for the coupler structures.

The analysis of coupler performance is based largely on two computational techniques. The first technique involves the use of the reciprocity theorem,[2] and the second involves a perturbation calculation using a transmission line analog.[3] Calculations based upon the full solution to the vector wave equation[4] are shown to agree with the approximate calculations.

The deleterious effects of waveguide thickness variations in the coupling region increases with increasing mode number for both the grating and the prism couplers. For the 25 μ thick wafer waveguide used in the early thin-film modulator, a thickness variation of about 8 μ over the input beam will reduce the input coupling for the TE_0 mode by a factor of 2; for TE_2 and TE_3 modes, thickness variations of 1.3 μ and 0.7 μ, respectively, will have the same deleterious effect. Because the coupling is relatively weak for the modulator waveguide, higher-order modes must be used for efficient coupling. Thickness variations on the order of 1 μ over a 1 to 2 mm^2 region of the input coupler were typically observed in the free-standing modulator waveguides.

Other structural imperfections that are particular to the grating coupler can degrade coupling efficiency. These imperfections include groove depth variations, variations in the profile of the grooves, and surface roughness which will cause scattering losses. The first two effects are treated in terms of variations in both effective waveguide thickness and the coupling parameter, and the last effect is calculated by using the results of an approximate diffraction model. Groove depth variations equal to half the average groove depth are considered. It is shown that for the modulator waveguide, the deleterious effects of groove geometry variations are minimized when the groove depth is approximately equal to the spatial extent of the evanescent field in the air region.

The partition of energy that results from multiple beams is analyzed for the grating coupling on a symmetric modulator waveguide. Because the lower diffracted beam can carry a significant amount of energy at both the input and the output gratings, the optical losses from multiple beam must be considered in the design of couplers. Variations in the partition of energy with the thickness of the coupler region are calculated, and the effects of such thickness variations on the intensity distribution of the output beam are discussed.

Experimental data with a free-standing waveguide are shown which indicate that structural imperfections in the coupler regions can be sufficiently large to influence

total coupling efficiency. These data include direct measurements of thickness variations in the waveguide, measurements of beam steering within the waveguide that occurs as a consequence of transverse thickness variations in the guide, and measurements of the intensity distribution of the output beam as a function of mode number. A direct comparison between coupling with a grating and coupling with a prism into the same waveguide shows that, for the modulator waveguide, the coupling efficiencies attainable with the prism are nearly an order-of-magnitude higher than those attainable with the grating. This striking difference is interpreted in terms of the analysis of structural imperfections.

This work was carried out as a joint effort between groups at the Polytechnic and at United Technologies Research Center.

Office of Naval Research
N00014-75-C-0421 D.W. Fradlin, P.K. Cheo, S.T. Peng and T. Tamir

National Science Foundation
ENG74-23908

REFERENCES

1. P.K. Cheo and M. Gilden, "Microwave Modulation of CO_2 Lasers in GaAs Optical Waveguides," Appl. Phys. Lett., 25, 272 (September 1974).
2. J.H. Harris, R.K. Winn and D.G. Dalgoutte, "Theory and Design of Periodic Couplers," Appl. Optics, 11, 234 (October 1972).
3. K. Handa, S.T. Peng and T. Tamir, "Improved Analysis of Dielectric Gratings," Appl. Phys., 5, 325 (1975).
4. S.T. Peng, T. Tamir and H.L. Bertoni, "Theory of Periodic Dielectric Waveguides," IEEE Trans., MTT-23, 123 (1975).

RF-SPUTTERING OF ALUMINUM-OXIDE FILMS FOR INTEGRATED OPTICS APPLICATIONS

L. Bergstein, E-W. Hu and M. Eschwei

Among all the existing vacuum deposition techniques, it seems that sputtering is one of the most attractive methods for the preparation of low-loss films for integrated optics applications. Despite the complex plasma processes associated with sputtering, the important sputtering parameters can be accurately controlled to yield reproducible results. Moreover, the deposition parameters such as gas content and sputtering powers can be readily adjusted over a wide range. This, combined with a relatively slow deposition rate, offers the possibility of deposition of films of variable refractive index. It is this last aspect which was the main motivation for the reported investigation of sputtered aluminum-oxide films.

Numerous works on thin films fabricated by sputtering techniques have been reported in the literature. We report in this work a systematic study of rf-sputtered aluminum-oxide films, emphasizing their potential applications in integrated optics. The choice of Al_2O_3 in this work is based on the following. First, Al_2O_3 seems to be a very attractive material for its proved versatile applications in conventional thin film optical devices area. It has been reported that thermally evaporated Al_2O_3 films are mechanically and chemically stable and transparent over a wide spectral range. Second, rf-sputtered Al_2O_3 films have been extensively investigated for microelectronic applications, mainly for their superior performance over SiO_2 films as a effective mask against dopants, alkaline ion diffusions, radiation damages, etc.[1] To our knowledge, no systematic and successful work on rf-sputtered Al_2O_3 film for integrated optics application has yet been reported. Efforts were made during the past year in this laboratory to obtain optical low-loss and reproducible rf-sputtered Al_2O_3 films. Using oxygen and/or argon, only moderate success was achieved. However, despite the lack of reproducibility and inability to fabricate very low loss films, the results were encouraging in the sense that the refractive index of the films did seem to vary with the sputtering parameters over a rather wide range.[2] We therefore decided to investigate the problem in a more detailed and systematic way.

We performed a series of runs using nitrogen and argon separately and in combination as the sputtering ion source. Since the sputtering processes take place at elevated temperatures, it was expected that the inclusion of oxygen would prevent the concentration of free aluminum (disassociated from the Al_2O_3 target at elevated temperatures) and thus enhance the quality of the oxide films. Contrary to our expectations, the results show that inclusion of oxygen in the sputtering chamber degrades the film. This is also indicated by Deitch et al.,[3] who found that oxygen is not the best atmosphere

for the deposition of oxides by sputtering. Similar observations were also reported recently[4] on some other oxide films prepared by dc-sputtering. Using only nitrogen and/or argon, we were able to obtain optical Al_2O_3 films of excellent quality. More importantly, by varying the sputtering parameters, we were able to vary the refractive indices of these films over a very wide range, from a low value of 1.48 at 6,328 Angstroms to a high value of 1.67.

A. Experimental Results

The experiment was carried out in a new Perkin-Elmer model 2400 sputtering system with a well trapped six-inch oil diffusion pump capable of yielding a vacuum better than 5×10^{-7} torr. Ultrahigh purity sputtering gases such as N_2, O_2 and Ar were admitted separately into the vacuum chamber and the pressures (of up to 40 microns) were accurately controlled and monitored. The hot-pressed Al_2O_3 target had a purity of 99.995%. Standard microscope slide glasses with refractive index of 1.5154 at 6328 Å were used as our basic substrates for most of the experiments. In order to obtain guided wave modes, fused quartz plates with a refractive index of 1.4567 were occasionally necessary when the refractive index of the films was lower than that of the microscope slides.

Immediately after proper cleaning, we transferred the substrates into the vacuum chamber and evacuated the system. The target was then pre-sputtered for at least 40 minutes in order to remove possible target surface contaminations. During the deposition, both the target and the base plate were water cooled. However, after about ten minutes the temperature of the substrates soon rose to and stabilized at about 200 to 350°C for sputtering powers of 100 to 500 Watts. A series of runs were performed by systematically changing the sputtering gas content and the sputtering rf-powers.

Our primary interest is to obtain low-loss films and to determine the dependence of the refractive index on the sputtering parameters. The prism coupler method[5] was used for measuring the refractive index. Typical film thickness was about two microns so that several modes were guided. The measurements for both TE and TM modes were carried out and the results showed no difference between the two cases. This assured the accuracy of the measurements and ruled out the possibility that the films are birefringent. Relative loss measurements were performed using the same setup described in Reference 2.

Figures 1 to 3 show some typical plots of the refractive indices of the films versus rf-sputtering powers with sputtering gas content (N_2 and/or Ar) as a parameter. (Since most films obtained with O_2 were lossy, they are not included in our results.) Figures 4 to 6 show the corresponding rates of deposition as a function of sputtering powers.

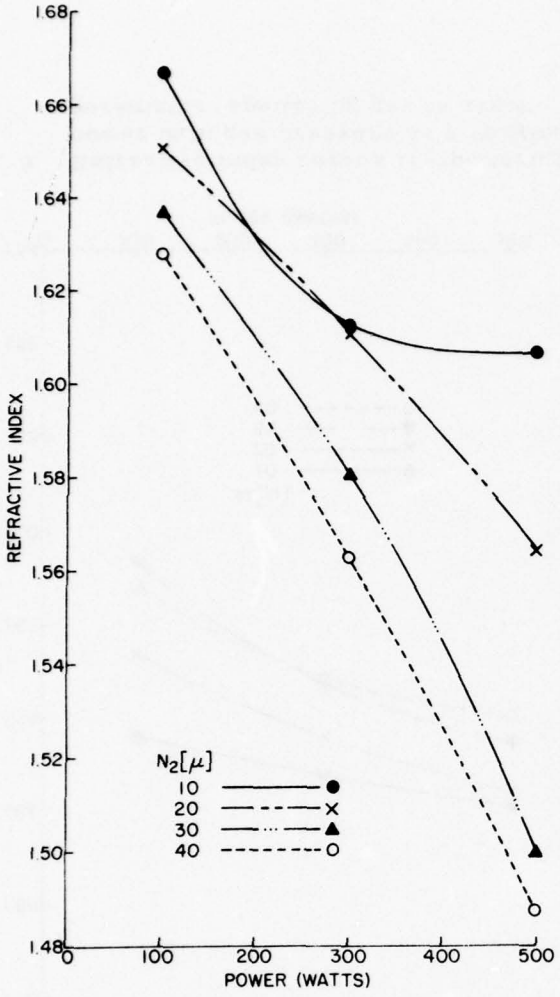


Fig. 1. Refractive index versus rf-sputtering power with gas pressure as a varying parameter; sputtering gas is nitrogen.

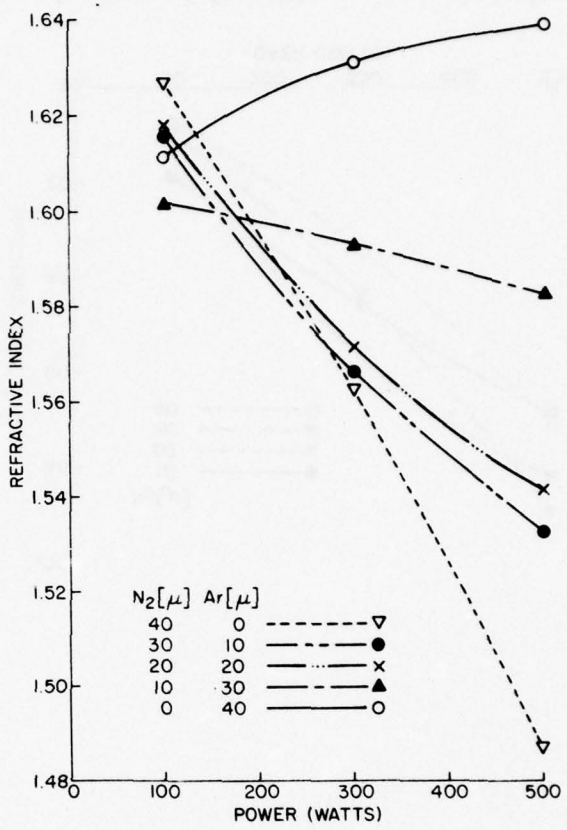


Fig. 2. Refractive index versus rf-sputtering power with gas pressure as a varying parameter; sputtering gas is a mixture of nitrogen and argon.

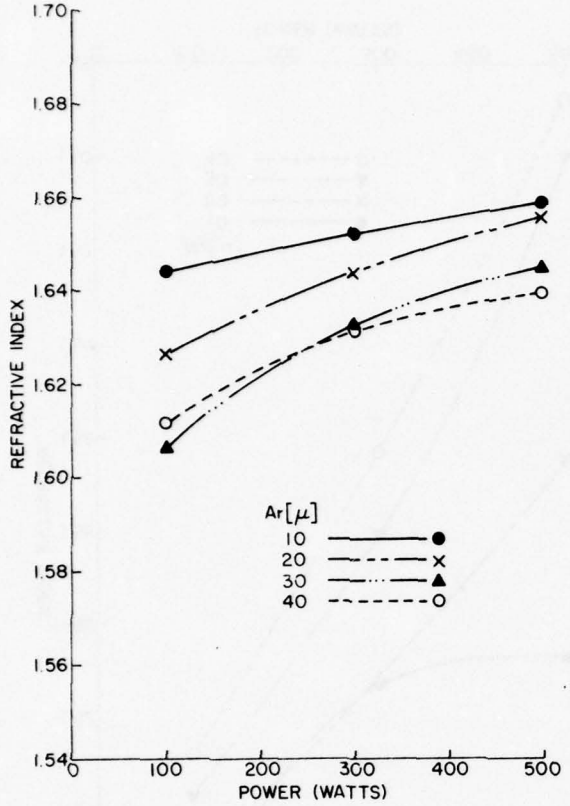


Fig. 3. Refractive index versus rf-sputtering power with gas pressure as a varying parameter; sputtering gas is argon.

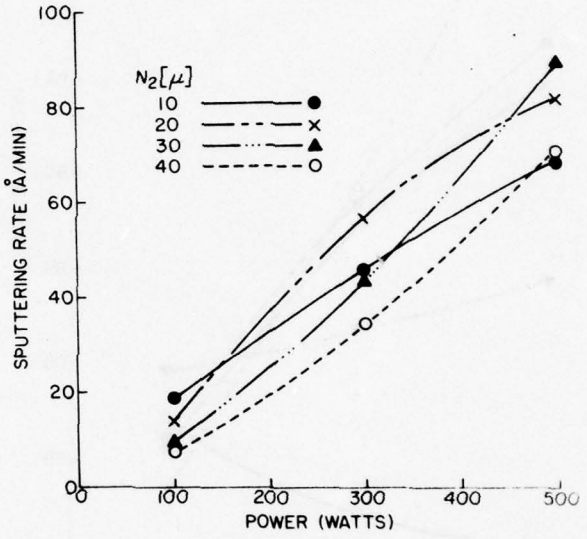


Fig. 4. Rate of deposition versus rf-sputtering power with gas pressure as a varying parameter; sputtering gas is nitrogen.

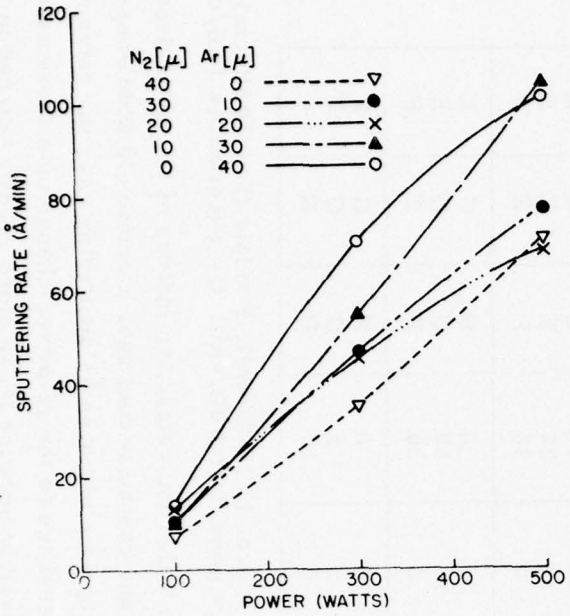


Fig. 5. Rate of deposition versus rf-sputtering power with gas pressure as a varying parameter; sputtering gas is a mixture of nitrogen and argon.

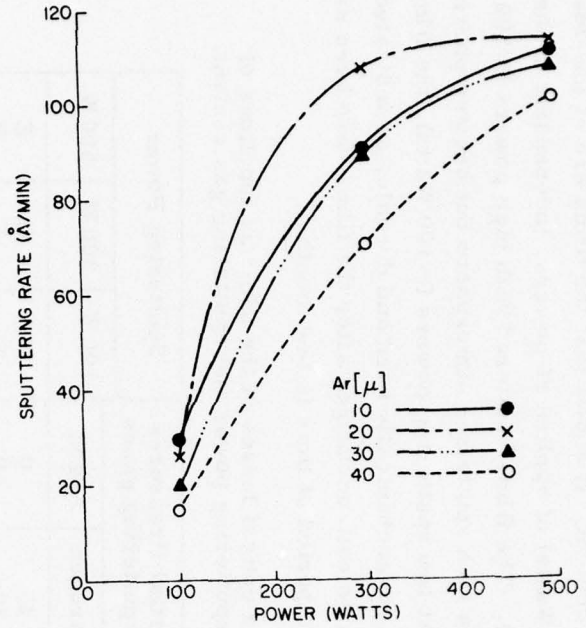


Fig. 6. Rate of deposition versus rf-sputtering power with gas pressure as a varying parameter; sputtering gas is argon.

Table I gives a convenient insight into the loss figures as a function of both rf-sputtering powers and sputtering gas content. It shows that the films are of low-loss with intermediate range (about 300 Watts) of applied rf-powers, independent of the sputtering gas content (N_2 and/or Ar). The films sputtered with high powers ($\cong$ 500 Watts) are generally good for nitrogen rich sputtering atmosphere but become poorer with increasing argon gas content. At low sputtering powers ($\cong$ 100 Watts), the films are in general lossy. All the films are mechanically hard and durable, as indicated by the fact that the refractive indices remain unchanged after the films have been exposed to the atmosphere for an extended period of time (a few months).

TABLE I. Relative optical losses of the films as functions of the rf-sputtering power and sputtering gas content.

Total Pressure	Partial Pressures of sputtering gases		Sputtering Power		
	Nitrogen	Argon	100 W	300 W	500 W
40 μ	40	0	F	G	E
	30	10	E	E	E
	20	20	G	F	P
	10	30	E	E	E
	0	40	P	F	P
30 μ	30 μ	0	F	G	G
	20	10	G	E	E
	10	20	E	E	E
	0	30	P	E	P
20 μ	20	0	P	G	E
	10	10	G	G	E
	0	20	P	E	P
10 μ	10	0	P	G	F
	5	5	G	G	E
	0	10	P	E	--

Relative Losses: P = poor (> 10 dB/cm), F = fair (5 to 10 dB/cm), G = good (2 to 5 dB/cm), E = excellent (< 1 dB/cm).

We note from Figs. 1 to 3 that the refractive indices of the films depend strongly on the deposition parameters and can be varied over a relatively wide range, from about 1.48 to 1.67. Moreover, using pure nitrogen as sputtering gas, the refractive indices of the films are much more sensitive to the rf-power change than those using pure argon as sputtering gas. This is possibly because N_2 is more readily trapped in the film than Ar. Efforts were made to relate the refractive indices to the film density. However, the results were not too conclusive and are not reported here. Some X-ray diffractometer and Laue-back reflection measurements were performed on a number of samples.

No crystalline structure was detected. Further work on the atomic level is probably necessary in order to gain a better understanding of the film structure and the sputtering processes.

It should be pointed out that the optical film losses can most probably be further reduced by using optically polished substrates, and a liquid nitrogen cold-trap to reduce diffusion pump back streaming.

B. Summary

We have successfully fabricated low-loss Al_2O_3 films with a wide range variation of refractive indices. This attractive feature, along with the mechanical and chemical durability of the films, makes the Al_2O_3 film a potential material for many optics and integrated optics applications. Examples are multilayer spectral filters and polarizers, polarization independent optical waveguides,[1] strip-line waveguiding structures, etc.

Joint Services Technical Advisory Committee
F44620-74-C-0056

L. Bergstein, E-W. Hu and M. Eschwei

REFERENCES

1. T.N. Kennedy, "Evaluating RF Sputtered Al_2O_3 for Microcircuit Fabrication," Electronic Packaging and Production, 136-141 (December 1974).

2. F.T. Stone and M. Eschwei, "Optical Properties of Sputtered Aluminum Oxides (Al_2O_3) Films for Application to Integrated Optics," Progress Report No. 39 to JSTAC, Polytech. Inst. of New York, Report No. R-452.39-74, 171-175 (1974).

3. R.H. Deitch, E.J. West, T.G. Giallorenzi and J.F. Weller, "Sputtered Thin Films for Integrated Optics," Appl. Optics, 13, 712-715 (1974).

4. S.J. Ingrey, W.D. Westwood, Y.C. Cheng and J. Wei, "Variable Refractive Index and Birefringent Waveguides by Sputtering Tantalum in O_2-N_2 Mixtures," Appl. Optics, 14, 2194-2198 (1975).

5. P.K. Tien, R. Ulrich and R.J. Martin, "Modes of Propagating Light Waves in Thin Deposited Semiconductor Films," Appl. Phys. Lett., 14, 291-294 (1969).

6. L. Bergstein, "A Polarization-Independent Optical Waveguide," Progress Report No. 36 to JSTAC, Polytech. Inst. of New York, Report No. R-452.36-71, 185-193 (1971).

COMPOSITION AND EQUILIBRIUM TOTAL PRESSURE OF BISMUTH VAPOR

W. T. Walter

During our efforts[1,2] to demonstrate efficient, pulsed laser action in the vapor of bismuth, it became apparent that a serious disagreement exists in the characterization of bismuth vapor as reported in several compilations. A large disparity exists between the most recent compilation of Nesmeyanov[3] and the prior ones of Hultgren et al.[4] and Stull and Sinke.[5] The disparity in dimer percentage is displayed in Fig. 1 by the three solid horizontal lines. The upper line line through Hultgren's points[4] indicates a dimer concentration of 50 to 60% in the pressure range of 1 mtorr to 10 torr. The middle line through Nesmeyanov's points[3] indicates a dimer concentration of about 20%, while the lower line through the points of Stull and Sinke[5] indicates a dimer concentration of approximately 10%.

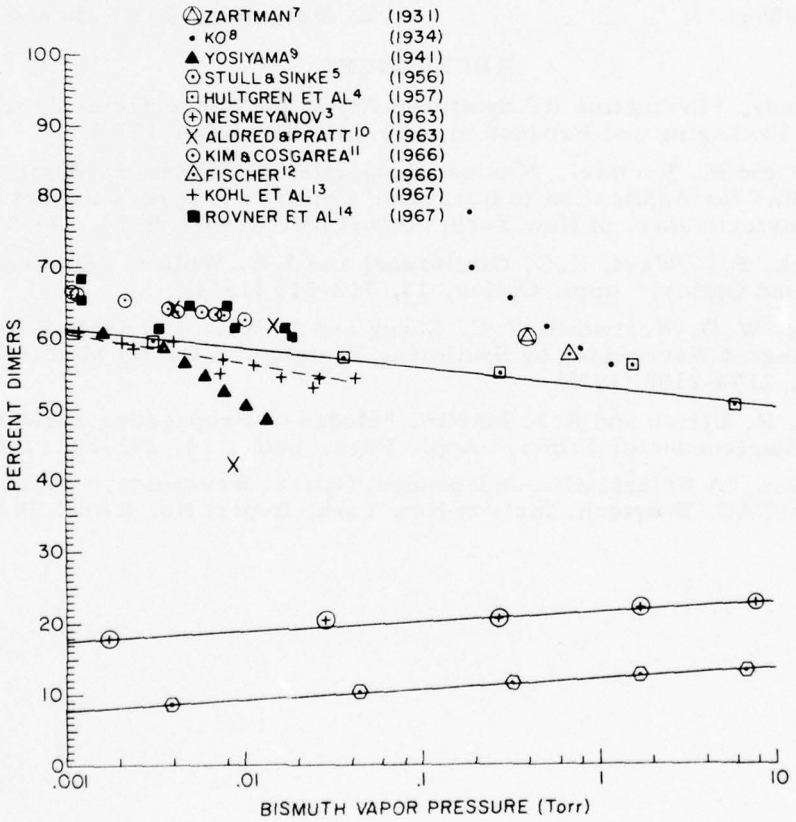


Fig. 1. Composition of bismuth vapor -- the percentage of dimers in the vapor as a function of the total equilibrium vapor pressure of bismuth.

The disparity in total bismuth vapor pressure is shown in Figure 2. The upper line through Nesmeyanov's points[3] indicates a total bismuth vapor pressure about ten times higher than either the compilation of Hultgren et al.[4] or Stull and Sinke[5] which are represented by the lower line.

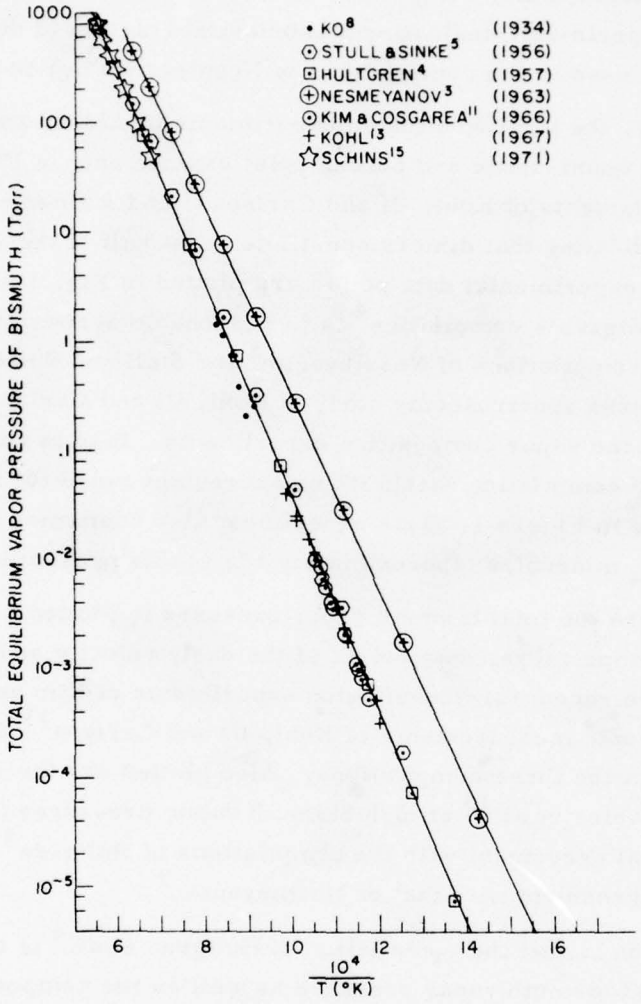


Fig. 2. Total equilibrium vapor pressure of bismuth as a function of the reciprocal of the absolute temperature.

To resolve these disparities, the early measurements were reexamined and compared with several recent measurements. On this basis, as described below, we conclude that the compilation of Hultgren et al.[4] best represents the actual situation in bismuth.

In 1928 Leu[6] carried out a Stern-Gerlach experiment using bismuth vapor and in 1931 Zartman[7] used a velocity analyzer on bismuth vapor. Both found that the vapor of bismuth consisted of about 50% Bi_2 molecules at total vapor pressures of 0.1 to 1 torr. Ko[8] corrected and extended these initial experiments and reported dimer concentrations of 60-70% at total vapor pressures of 0.1 to 1 torr. In 1941 Yosiyama[9] using a torsion-effusion method obtained dimer compositions of the bismuth vapor of 40-50% at 1 to 10 mtorr. These experiments, indicating a substantial fraction of dimers in the vapor of bismuth, were not used in the compilations of Nesmeyanov[3] or Stull and Sinke.[5]

More recently, the torsion-effusion experiments of Aldred and Pratt[10] and Kim and Cosgarea,[11] the quasi-static and boiling point experiments of Fischer[12] and the mass spectrometric experiments of Kohl, Ul and Carlson[13] and Rovner et al.[14] all confirm the earlier work indicating that dimers constitute about half of the vapor of bismuth. These more recent experimental data points are plotted in Fig. 1 along with the earlier measurements. Hultgren's compilation[4] is in reasonable agreement with all of the data in Fig. 1 while the compilations of Nesmeyanov[3] and Stull and Sinke[5] are not and must be rejected. The mass spectroscopic study of Kohl, Ul and Carlson[13] appears to be the most reliable of all the vapor composition experiments. It is in good agreement with the earlier Hultgren compilation within its measurement range (0.3-42 mtorr) as indicated by the dotted line in Figure 1. This experiment also confirmed the presence of a small amount of Bi_4 molecules (approximately 1%) in this pressure region.

In Fig. 2 where the total bismuth vapor pressure is plotted as a function of the reciprocal of the temperature, data points of the early velocity analyzer measurements of Ko,[8] and the more recent torsion-effusion experiments of Kim and Cosgarea[11] and the mass spectroscopic measurements of Kohl, Ul and Carlson[13] are displayed in addition to points from the three compilations. Also plotted are the recent heat-pipe measurements of Schins et al.[15] at high bismuth vapor pressures (> 49 torr). All of this data is in substantial agreement with the compilations of Hultgren[4] and Stull and Sinke[5] and in serious disagreement with that of Nesmeyanov.[3]

Our conclusion is that the compilation of Hultgren et al.[4] is the best characterization of both the total bismuth vapor pressure as well as the composition of the vapor in terms of monomers and dimers. Therefore the values of Hultgren have been utilized in our double-pulse discharge examination[2] of bismuth vapor for efficient, pulsed laser action at 4722 Å.

Office of Naval Research
N00014-67-A-0438-0015

W.T. Walter

ARPA under Office of Naval Research
N00014-67-A-0438-0017

Joint Services Technical Advisory Committee
F44620-74-C-0056

REFERENCES

1. K. Park, N. Solimene and W.T. Walter, "Investigation of Cyclic Laser Action in Bismuth Vapor," Progress Report No. 38 to JSTAC, Polytech. Inst. of New York, Report No. R-452.38-73 (1973).
2. W.T. Walter, "Double-Pulse Discharge Examination of Bismuth Vapor for Efficient, Pulsed Laser Action at 4722 Å," Progress Report No. 41 to JSTAC, Polytech. Inst. of New York, Report No. R-452.41-76 (1976).
3. A.N. Nesmeyanov, "Vapor Pressure of the Chemical Elements," R. Gary, Ed. (Netherlands: Elsevier, 1963).
4. R. Hultgren, R.L. Orr, P.D. Anderson and K.K. Kelley, "Selected Values of Thermodynamic Properties of Metals and Alloys," (New York: John Wiley and Sons, 1963).
5. D.R. Stull and G.C. Sinke, "Thermodynamic Properties of the Elements," Advances in Chemistry Series No. 18, American Chemical Society, Washington, D.C. (1956).
6. A. Leu, "Untersuchungen an Wismut nach der magnetischen Molekularstrahlmethode," Zeitschrift für Physik 49, 498 (1928).
7. I.F. Zartman, "A Direct Measurement of Molecular Velocities," Phys. Rev. 37, 383 (1931).
8. C.C. Ko, "The Heat of Dissociation of Bi_2 Determined by the Method of Molecular Beams," J. Franklin Institute 217, 173 (1934).
9. M. Yosiyama, J. Chem. Soc. Japan 62, 204 (1941). (The values used in this report were those quoted in Table 252, page 303 of Ref. 3 above.)
10. A.T. Aldred and J.N. Pratt, "Vapor Pressure of Liquid Bismuth," J. Chem. Phys. 38, 1085 (1963).
11. J.H. Kim and A. Cosgarea, Jr., "Study of the Vapors of Liquid Lead and Bismuth," J. Chem. Phys. 44, 806 (1966).
12. A.K. Fischer, "Vapor Pressure of Bismuth," J. Chem. Phys. 45, 375 (1966).
13. F.J. Kohl, O.M. Ul and K.D. Carlson, "Cross Sections for Electron-Impact Fragmentation and Dissociation Energies of the Dimer and Tetramer of Bismuth," J. Chem. Phys. 47, 2667 (1967).
14. L. Rovner, A. Drowart and J. Drowart, "Mass Spectrometric Determination of Dissociation Energies of Molecules Bi_2, Bi_3, Bi_4 and BiPb," Trans. Faraday Society 63, 2906 (1967).
15. H.E.J. Schins, R.W.M. van Wijk and B. Dorpema, "The Heat-Pipe Boiling-Point Method and the Vapor Pressure of Twelve Metallic Elements in the Range $10\text{-}10^4$ Torr," Zeitschridt für Metallkunde 62, 330 (1971).

DOUBLE-PULSE DISCHARGE EXAMINATION OF BISMUTH VAPOR FOR EFFICIENT, PULSED LASER ACTION AT 4722 Å

W. T. Walter and K. Park

The kinetics of electron collisions in pulsed electrical discharges in atomic vapors with suitable energy level structures can produce efficient, pulsed laser action.[1] This has been demonstrated in the vapors of Pb,[2] Mn,[3] Cu,[4] Au,[5] Ca,[6] Sr,[6] and Ba.[7] Most experimental effort thus far has been applied to the copper vapor laser. Peak powers of 170 kW and average powers of 15 W have been generated at 5105 Å in copper vapor with a 1% overall electrical efficiency.[8]

Similar laser action is possible in other elements as indicated in Table IV of Reference 1. The possibility at 4722 Å in bismuth is of particular interest because it would be close to the wavelength of maximum transmission through water. However, in spite of attempts in a number of different laboratories, laser action has not, to our knowledge, been observed at 4722 Å in the vapor of bismuth.

The vapor of bismuth contains a substantial percentage of dimers while the vapors of copper and lead do not. Figure 1 compares the dimer concentration in the vapors of bismuth,[9,10] copper[11] and lead[11] in the pressure range 0.1 mtorr to 100 torr. This is a broader pressure range by a factor of at least ten at each end, than the copper and lead vapor lasers have operated in thus far. The dimer concentration in bismuth vapor is greater than 50% at pressures below 10 torr while the dimer concentration in copper vapor is less than 1% and in lead vapor less than 0.1% in a similar pressure region. At 1 torr of total metallic vapor pressure, for example, where operation of both the copper and lead vapor lasers is very strong, the dimer concentration in bismuth vapor is 53% while it is only 0.4% in copper vapor and 0.06% in the vapor of lead.

The presence of bismuth dimers could adversely affect the possibility of laser action at 4722 Å in several ways:

(1) directly by absorption at 4722 Å

(2) by absorption of the 3067 Å bismuth resonance line which will not only reduce radiation trapping but can produce dissociation of the Bi_2 dimers leaving one Bi atom in the $6p^3\,{}^2D^o_{3/2}$ metastable proposed lower laser level,[12,13]

(3) by means of collisional processes which interfere with the excitation process, such as a modification of the electron distribution, or which quench the upper resonance level.

To determine whether the presence of dimers in the vapor of bismuth could be the cause for non-observance of laser action at 4722 Å, we experimentally examined two methods of dissociating the dimers and tested the resulting vapor for laser action: 1) thermal dissociation and 2) discharge dissociation.

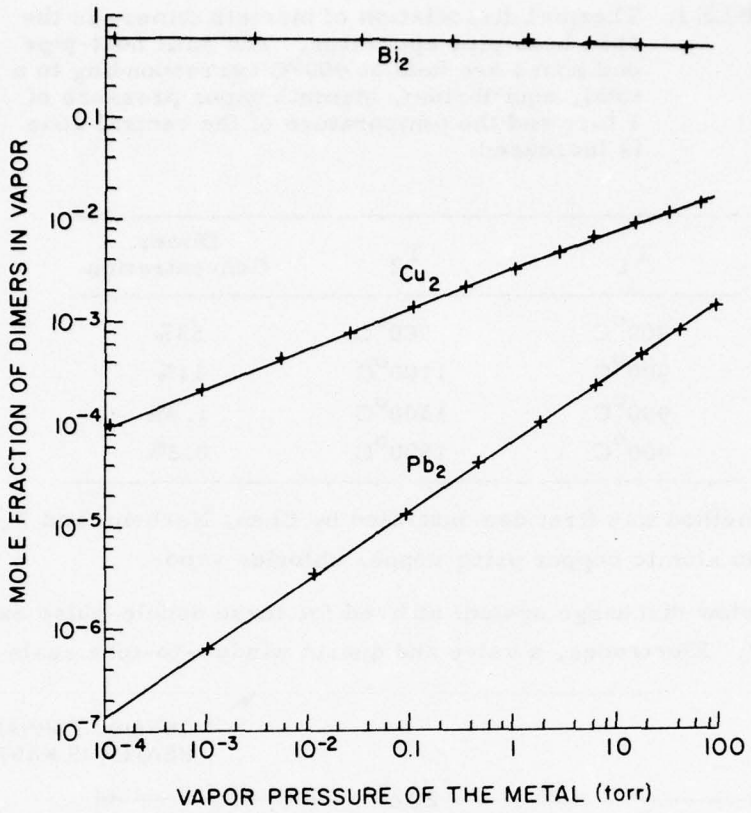


Fig. 1. Comparison of the mole fraction of dimers in the vapors of bismuth, copper and lead.

The split heat-pipe apparatus previously described[14] was used to create two distinct temperature zones in a furnace and thermally dissociate the Bi_2 molecules in the hotter central zone as indicated in Table I. The temperature of the split heat-pipe end zones was held at 900°C so that the pressure throughout the apparatus remained constant at 1 torr. The values calculated in Table I are based on Hultgren's values both of dimer concentration and of total bismuth vapor pressure according to our critical evaluation[9] of the experimental data.

When thermal dissociation was experimentally tested, establishment of two distinct temperature zones was evident. The temperature along the 6" long tungsten-mesh split heat-pipe was examined by means of an optical pyrometer. The temperature indicated along the inside of the mesh tube was uniform confirming the wicking action of liquid bismuth. The split heat-pipe also served as electrodes for the electric discharge. No evidence of laser action at 4722 Å was observed, however.

A double-pulse discharge system was then utilized to dissociate the bismuth molecules -- the first pulse to dissociate the Bi_2 dimers and the second to excite the

TABLE I. Thermal dissociation of bismuth dimers in the split heat-pipe apparatus. The split heat-pipe end zones are held at 900°C corresponding to a total, equilibrium, bismuth vapor pressure of 1 torr and the temperature of the central zone is increased.

T_1	T_2	Dimer Concentration
900°C	900°C	53%
900°C	1100°C	11%
900°C	1300°C	1.4%
900°C	1500°C	0.3%

Bi atoms. This method was first demonstrated by Chen, Nerheim and Russell[15] to obtain laser action in atomic copper using copper chloride vapor.

The hot-window discharge system utilized for these double-pulse experiments is shown in Figure 2. Electrodes, a valve and quartz window-to-tube seals all capable

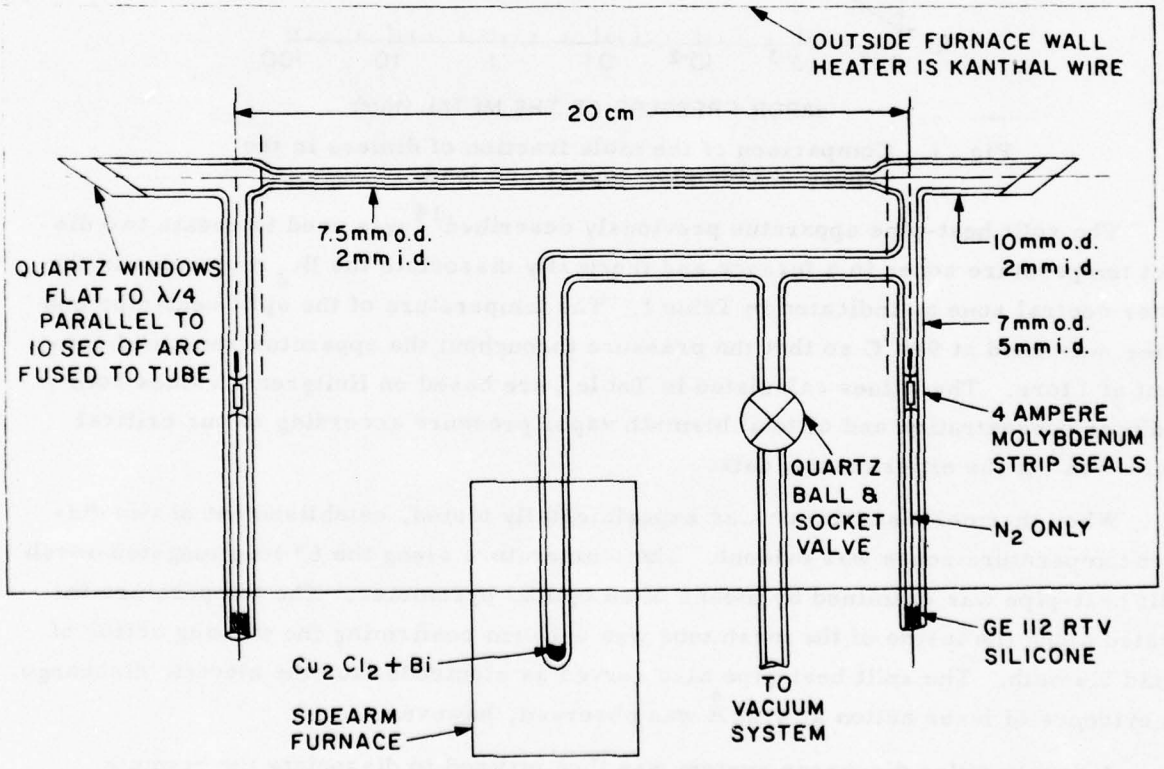


Fig. 2. Hot-window quartz-tube laser-discharge apparatus capable of sustained operation at temperatures up to 1000°C for copper halide and bismuth dimer vapors.

of withstanding temperatures as high as 1000°C were developed to construct a system in which the entire laser tube can be operated at temperatures up to the 1000°C devitrification limit of quartz. The entire laser discharge tube can operate then, under equilibrium temperature and vapor density conditions; the the operational lifetime is no longer limited by migration of the active material out of the hot zone.

The quartz laser discharge tube was connected to a sidearm, as indicated in Fig. 2, which contained 0.2 gms of copper chloride and 1.25 gms of bismuth shot. The sidearm was placed in a second furnace to independently control its temperature and thereby the vapor pressure of its contents. The optical cavity consisted of an aluminized concave mirror (172.4 cm radius) and a flat beamsplitter (63% transmission) separated by 120 cm.

The double-pulse discharge system is shown in Figure 3. Two Sprague 3900 pF

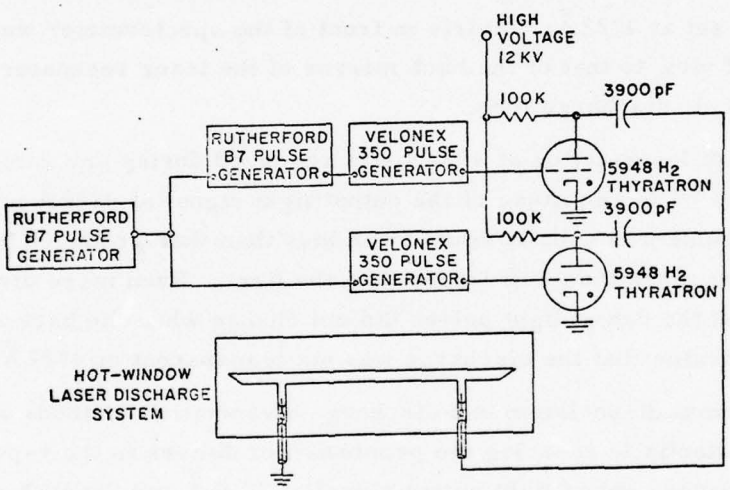


Fig. 3. Double-pulse excitation system for copper halide and bismuth dimer vapors.

ceramic capacitators charged to 12 kV were independently discharged through the laser tube by separate hydrogen thyratrons. The pulse generators provide an adjustable delay time between the two pulses as well as an adjustable repetition rate of the double-pulse signal.

The small amount of copper chloride was included in the sidearm along with bismuth to demonstrate the efficacy of double-pulse excitation in our experimental system. When the temperature of the main and sidearm furnaces was raised to 500°C, laser action was observed on the 5105 Å and 5782 Å atomic copper transitions. Optimum delay

time between pulses was 28-30 μsec. The pressure range in which the laser would operate was 3-5 torr of argon with the optimum being 3.8 torr. Peak current amplitudes were 13 and 15 amperes respectively for the two pulses, corresponding to peak current densities of 400 to 500 A/cm^2. The copper vapor laser action was also used to optimize the optical cavity alignment.

When the temperature of the furnaces were raised above 700°C, the copper chloride diffused through the open quartz ball and socket valve and bismuth took over the discharge. The furnace temperatures were gradually taken as high as 970°C corresponding to an equilibrium total bismuth vapor pressure of 2.7 torr. Helium and argon were used as buffer gases and tested throughout a pressure range of 0-500 torr. The peak current amplitudes were operated up to 20 amperes (600 A/cm^2). The time delay between pulses was usually set at 30 μsec. When the delay was varied, the output light signal at 4722 Å produced by the second pulse never increased from its value at 30 μsec. The output light pulses were examined by means of a Jarrell-Ash 0.5m spectrometer which was usually set at 4722 Å. An iris in front of the spectrometer was adjusted to narrow the field of view to that of the back mirror of the laser resonator as viewed through the 2mm i.d. discharge tube.

No evidence of laser action at 4722 Å was observed during any variation of experimental parameters. The amplitude of the output light signal at 4722 Å produced by the second discharge pulse was usually equal to or less than that produced by the first pulse. It was never more than 20% larger than the first. Even more disheartening was the observation that the output light pulses did not change when the back mirror was blocked. This indicates that the discharge was not transparent at 4722 Å.

Both the thermal dissociation and discharge dissociation methods were successful, we believe, in substantially reducing the proportion of dimers in the vapor of bismuth. The absence of enhancement of light output signal at 4722 Å and the lack of transparency of the discharge at the proposed laser transition in spite of an estimated reduction in dimer mole fraction in the vapor by 100 suggests that other processes are more important in populating the $^2D_{3/2}$ proposed lower laser level. We conclude, therefore, that the presence of dimers in the vapor of bismuth is not the dominant explanation for the absence of laser action at 4722 Å.

Office of Naval Research
N00014-67-A-0438-0015

W. T. Walter

ARPA under Office of Naval Research Contract
N00014-67-A-0438-0017

Joint Services Technical Advisory Committee
F44620-74-C-0056

REFERENCES

1. W.T. Walter, N. Solimene, M. Piltch and G. Gould, "Efficient Pulsed Gas Discharge Lasers," IEEE J. Quantum Electron., QE-2, 474 (1966).
2. G.R. Fowles and W.T. Silfvast, "High-Gain Laser Transition in Lead Vapor," Appl. Phys. Lett., 6, 236 (1965).
3. M. Piltch, W.T. Walter, N. Solimene, G. Gould and W.R. Bennett, Jr., "Pulsed Laser Transitions in Manganese Vapor," Appl. Phys. Lett., 7, 309 (1965).
4. W.T. Walter, M. Piltch, N. Solimene and G. Gould, "Pulsed-Laser Action in Atomic Copper Vapor," Bull. Am. Phys. Soc., 11, 113 (1966).
5. W.T. Walter, "Metal Vapor Lasers," IEEE J. Quantum Electron., QE-4, 355 (1968).
6. J.S. Deech and J.H. Sanders, "New Self-Terminating Laser Transitions in Calcium and Strontium," IEEE J. Quantum Electron., QE-4, 474 (1968).
7. P. Cahuzac, "Raies Laser Infrarouges dans les Vapeurs de Terres Rares et d'Alcalino-Terreux," J. Phys. (Paris) 32, 499 (1971).
8. A.A. Isaev, M.A. Kazaryan and G.G. Petrash, "Effective Pulsed Copper-Vapor Laser with High Average Generation Power," (JETP Lett. 16, 27 (1972)) ZhETF Pis. Red. 16, 40 (1972).
9. W.T. Walter, "Composition and Equilibrium Total Pressure of Bismuth Vapor," Progress Report No. 41 to JSTAC, Polytech. Inst. of New York, Report No. R-452.41-76 (1976).
10. R. Hultgren, R.L. Orr, P.D. Anderson and K.K. Kelley, "Selected Values of Thermodynamic Properties of Metals and Alloys," (John Wiley and Sons, New York, 1963).
11. *D.R. Stull and H. Prophet*, "JANAF Thermochemical Tables," Second Edition, Nat. Stand. Ref. Data Ser., Nat. Bur. Stand. (U.S.), 37 (U.S. Govt. Printing Office, Washington, D.C., June 1971).
12. G.M. Almy and F.M. Sparks, "The Absorption Spectrum of Diatomic Bismuth," Phys. Rev., 44, 365 (1933).
13. G. Gerber and H.P. Broida, "Electronic States and Molecular Constants of Bi_2," J. Chem. Phys., 64, 3423 (1976).
14. K. Park, N. Solimene and W.T. Walter, "Investigation of Cyclic Laser Action in Bismuth Vapor," Progress Report No. 38 to JSTAC, Polytech. Inst. of New York, Report No. R-452.38-73 (1973).
15. C.J. Chen, N.M. Nerheim and G.R. Russell, "Double-Discharge Copper Vapor Laser with Copper Chloride as a Lasant," Appl. Phys. Lett., 23, 514 (1973).

SPONTANEOUS-EMISSION TRANSITION PROBABILITY OF THE 4722 Å ATOMIC BISMUTH LINE

W. T. Walter and N. Solimene

One of the criteria for efficient, pulsed atomic-vapor lasers[1] is that the risetime of the excitation current pulse must be faster than the reciprocal of the transition probability, $A_{u\ell}$, of the proposed laser line for laser action to be achieved in a three level energy level structure as indicated in Figure 1. If this criterion is not satisfied, spontaneous radiation will drain the upper laser level and fill the metastable lower laser level, thereby preventing the establishment of a population inversion. The total upper level lifetime $(A_{u\ell} + A_{ug})^{-1} \approx A_{u\ell}^{-1}$, because the resonance radiation is well trapped at the operating temperature for pulsed laser action. For example, in a 1 cm diam tube at 1100°K containing 0.1 torr of atomic bismuth vapor pressure,[2,3] the radiation-trapped transition probability[4] of the 3067 Å resonance line has been reduced from 2×10^8 to 1×10^5 sec^{-1}; and the atomic bismuth vapor presence for optimum peak laser output is expected to be > 0.1 torr.

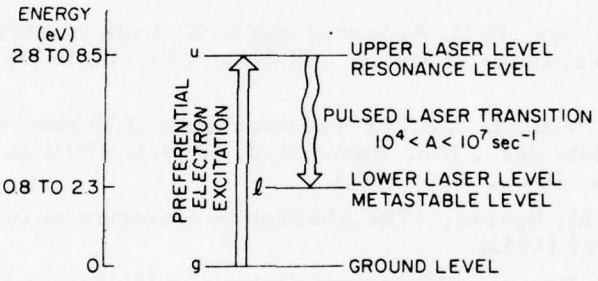


Fig. 1. Efficient pulsed gas-discharge laser in an atomic vapor.

In our examination[5,6] of efficient, pulsed laser action at 4722 Å in the atomic vapor of bismuth, the only measurement we have found for the 4722 Å spontaneous-emission transition lifetime is the 111 nsec value of Corliss and Bozman.[7] Substantial errors, some as large as a factor of 20, have been reported in the values of transition probabilities determined by Corliss and Bozman[7] from the intensity measurements of Meggers, Corliss and Scribner.[8] Therefore one is reluctant to rely very strongly on these values. Because accurate values of atomic transition probabilities are important in the evaluation of new laser systems, we developed a procedure[9] to obtain improved values of the Corliss-Bozman transition probabilities. When Fig. 1 of Ref. 9 is used with the bismuth parameters (E_u = 32588 cm^{-1}, $I = 60 \times 10^3$, λ = 4722 Å and U = 4.3), an improved Corliss-Bozman transition lifetime of 140 nsec is obtained for the 4722 Å line.

Although apparently no other direct measurements have been carried out on the 4722 Å bismuth line, a number of measurements have been made on the 3067 Å resonance

line and of the lifetime of the $6p^2\,7s\,{}^4P_{1/2}$ resonance level which is the proposed upper laser level. These measurements, which are listed in Table I, can be used with the branching ratio to obtain a better value for the transition lifetime of the 4722 Å bismuth line.

TABLE I. Lifetime of $6p^2\,7s\,{}^4P_{1/2}$ bismuth resonance level.

Measurement	Method	Lifetime (nsec)
Corliss and Bozman[7]	Emission	2.8
Lvov[10]	Absorption	8.9 ± 1.8*
Rice and Ragone[11]	Absorption	$5.3^{+2.3}_{-1.7}$*
Cunningham and Link[12]	Phase Shift	5.9 ± 0.2
Svanberg[13]	Hanle Effect	4.75 ± 0.18
Anderson, et al.[14]	Beam Foil	4.7 ± 1.0

*As corrected in Ref. 12 by using the Corliss-Bozman value of 0.97 for the branching ratio into the 3067 Å resonance line.

Intermediate Coupling SCF Calculation	Type of Exchange Approximation	Form	Lifetime
Holmgren[15]	Hartree-Slater	Dipole Length	5.7 nsec
		Dipole Velocity	2.6 nsec
Holmgren[15]	OHFS-Lingren-Rosen	Dipole Length	4.1 nsec
		Dipole Velocity	7.1 nsec
Kunisz and Migdalek[16]	Lingren		5.7 nsec

Table I reveals that all of the other experimental measurements of the $6p^2\,7s\,{}^4P_{1/2}$ resonance level lifetime are longer than Corliss and Bozman's 2.8 nsec value. The most accurate value is probably Svanberg's Hanle-effect measurement of 4.75 nsec.

Recently intermediate coupling calculations have been carried out for $6p^2\,7s \rightarrow 6p^3$ transitions in bismuth.[15,16] The calculated lifetime values are listed in the lower part of Table I. The transition probabilities calculated depend on the local exchange approximation as well as on the dipole length or dipole velocity forms of the transition probability operator since the self-consistent field wavefunctions are not exact. The difference between the dipole length and dipole velocity calculations can be viewed as an

indication of the inexactness of the wavefunctions. The calculations indicate that the lifetime of the $6p^2\,7s\,{}^4P_{1/2}$ bismuth resonance level is approximately 5 nsec which is consistent with the best experimental values and is substantially longer than the Corliss-Bozman value.

The branching ratio, $A_{u\ell} / \sum_{\text{all } j} A_{uj} = A_{u\ell}\,\tau_u$, may be used to obtain a particular transition lifetime from the level lifetime, τ_u. The only experimental values available to determine branching ratios in bismuth are from Corliss and Bozman.[7] These ratios, however, should be more accurate than the value for an individual transition probability since effects other than radiation trapping, such as uncertainty in the degree of ionization or non-uniformity of the arc, are expected to be the dominant sources of error and should substantially cancel out in the ratio.

Branching ratios into the 4722 Å bismuth transition are also available from the intermediate coupling calculations.[15,16] These values are in substantial agreement with the Corliss-Bozman value as indicated in Table II.

TABLE II. Branching ratio into the 4722 Å bismuth transition.

Source	Method	Branching Ratio
Corliss-Bozman[7]	Emission Measurement	0.025
Holmgren[15]	Calculation - OHFS -	
	dipole length	0.020
	dipole velocity	0.047
Kunisz-Migdalek[16]	Calculation	0.026

Using the Corliss-Bozman branching ratio of .025 for the 4722 Å $^4P_{1/2} \rightarrow {}^2D_{3/2}$ line, we may deduce a 190 nsec transition lifetime from Svanberg's 4.75 nsec $^4P_{1/2}$ level lifetime. This is substantially longer than the 111 nsec Corliss-Bozman value and also longer than the improved value of 140 nsec.

All indications are, therefore, that the transition lifetime of the 4722 Å proposed laser transition in bismuth is significantly greater than 100 nsec; the best value being 190 nsec. This corresponds to a spontaneous-emission transition probability for the 4722 Å bismuth line of 5.3×10^6 sec^{-1}. Excitation pulse risetimes in our various metal vapor discharge tubes have been 50 to 100 nsec. During the course of our experimental investigation of pulsed discharges in bismuth vapor,[5,6] we have produced excitation current risetimes as short as ~25 nsec in a radial discharge between a wire at the center and a cylindrical electrode at the wall of the discharge tube. Since excitation

pulse risetimes as short as ~25 nsec have been produced and since laser action has been produced on transitions with similar or shorter transition lifetimes in Mn and Pb; it appears to be very unlikely that an insufficiently fast risetime of the current excitation pulse can be the explanation for the absence of laser action at 4722 Å in bismuth vapor.

Office of Naval Research
N00014-67-A-0438-0015

W.T. Walter

ARPA under Office of Naval Research Contract
N00014-67-A-0438-0017

Joint Services Technical Advisory Committee
F44620-74-C-0056

REFERENCES

1. W.T. Walter, N. Solimene, M. Piltch and G. Gould, "Efficient Pulsed Gas Discharge Lasers," IEEE J. Quantum Electron., QE-2, 474 (1966).
2. W.T. Walter, "Composition and Equilibrium Total Pressure of Bismuth Vapor," Progress Report No. 41 to JSTAC, Polytech. Inst. of New York, Report No. R-452.41-76 (1976).
3. R. Hultgren, R.L. Orr, P.D. Anderson and K.K. Kelley, "Selected Values of Thermodynamic Properties of Metals and Alloys," (John Wiley and Sons, New York, 1963).
4. D. Alpert, A.O. McCoubrey and T. Holstein, "Imprisonment of Resonance Radiation in Mercury Vapor," Phys. Rev., 76, 1257 (1949). A factor of τ is missing from the formula given in footnote 7 of this reference.
5. K. Park, N. Solimene and W.T. Walter, "Investigation of Cyclic Laser Action in Bismuth Vapor," Progress Report No. 38 to JSTAC, Polytech. Inst. of New York, Report No. R-452.38-73 (1973).
6. W.T. Walter and K. Park, "Double-Pulse Discharge Examination of Bismuth Vapor for Efficient, Pulsed Laser Action at 4722 Å," Progress Report No. 41 to JSTAC, Polytech. Inst. of New York, Report No. R-452.41-76 (1976).
7. C.H. Corliss and W.R. Bozman, "Experimental Transition Probabilities for Spectral Lines of Seventy Elements," Natl. Bur. Stand. Monograph No. 53 (U.S. Govt. Printing Office, Washington, D.C., 1962).
8. W.F. Meggers, C.H. Corliss and B.F. Scribner, "Tables of Spectral Line Intensities," Natl. Bur. Stand. Monograph No. 32 (U.S. Govt. Printing Office, Washington, D.C., 1961).
9. W.T. Walter and N. Solimene, "Improved Values of Atomic Transition Probabilities," Progress Report No. 40 to JSTAC, Polytech. Inst. of New York, Report No. R-452.40-75 (1975).
10. B.V. Lvov, "Determination of the Absolute Values of Oscillator Strengths by Combined Measurement of the Total and Linear Absorption of Vapor Layers in a Graphite Cell," Optics and Spectroscopy, 19, 282 (1965).
11. P.A. Rice and D.V. Ragone, "Simultaneous Determination of f Values and Vapor Pressures from Optical Absorption Measurements," J. Chem. Phys., 42, 701 (1965).
12. P.T. Cunningham and J.K. Link, "Measurement of Lifetimes of Excited States of Na, Tl, In, Ga, Cu, Ag, Pb, and Bi by the Phase-Shift Method," J. Opt. Soc. Am., 57, 1000 (1967).

13. S. Svanberg, Thesis, University of Gothenburg (May 1972), as quoted in Reference 14.

14. T. Andersen, O.H. Madsen and G. Sørensen, "Radiative Lifetimes in SnI and BiI," J. Opt. Soc. Am., 62, 1118 (1972).

15. L. Holmgren, "Theoretically Calculated Transition Probabilities and Lifetimes for the First Excited Configuration $np^2(n+1)s$ in the Neutral As, Sb, and Bi Atoms," Physica Scripta, 11, 15 (1975).

16. M.D. Kunisz and J. Migdalek, "Theoretical Values of Certain Transitions of Type $6p^3$-$6p^2 7s$ in the Bismuth Atomic Spectrum in Intermediate Coupling," Acta Physica Polonica, A47, 231 (1975).

APPLICATION OF THE HEISENBERG PICTURE TO THE AVERAGED DESCRIPTION OF THE PROPAGATION OF OPTICAL BEAMS

M.C. Newstein and D. Ramakrishnan

A large class of problems describing the propagation of optical signals is described by the quasi-optic equation

$$i\lambda \frac{\partial}{\partial z} \mathcal{E} = \left[\frac{(i\lambda \nabla_T)^2}{2} + V \right] \mathcal{E} \tag{1}$$

where $\mathcal{E}$ is the slowly varying complex envelope of an optical signal, ∇_T is the transverse laplacian and V is proportional to the susceptibility of the medium, $\lambda = \lambda/2\pi$. Formal integrals such as

$$\langle z|F|z\rangle = \iint dxdy\, \mathcal{E}^*(x,y,z)\, F(x,y,p_x,p_y;z)\, \mathcal{E}(x,y;z) \tag{2}$$

where

$$p_x = -i\lambda \partial_x \,, \quad p_y = -i\lambda \partial_y$$

may be used as concise representations of important field characteristics.

The optics problem of finding the dependence on z of the weighted operator function $\langle F \rangle$ defined by Eq. (2) is analogous to the quantum mechanics problem of determining the evolution, in time, of the expectation value of a dynamical operator F. The states of a quantum system correspond to vectors, $|\psi\rangle$ in a Hilbert space. Dynamical variables, such as position, x, and momentum p_x, correspond to linear transformations in this space.[1]

Alternative pictures are available for the interpretation of the dynamical evolution of the wave function.[2] In the Schrodinger picture the state vectors are visualized as evolving in time through a fixed coordinate frame determined by the eigenvectors of time independent operators. The quantum expectation value of Schrodinger operator[3] $F(x,p_x;t)$, in the Schrodinger state $|\phi(t)\rangle$ is given in the x representation, by

$$\langle\psi(t)|F(x,p_x,t)|\psi(t)\rangle = \int dx\psi^*(x,t)\, F(x,-i\hbar\frac{\partial}{\partial x},t)\,\psi(x,t) \tag{3}$$

In the Heisenberg picture the state vectors are visualized as fixed in time, but the operators and their eigen-vectors are time dependent. The quantum expectation value of the Schrodinger operator $F(x,p_x;t)$ in the state represented by the vector $|\psi(t)\rangle$ equal to the quantum expectation value of the Heisenberg operator $F(x(t),p_x(t);t)$ in the initial state represented by the vector $|\psi(t_0)\rangle$, i.e.,

$$\langle\psi(t)|F(x,p_x;t)|\psi(t)\rangle = \langle\psi(t_0)|F(x(t),p_x(t);t)|\psi(t_0)\rangle \tag{4}$$

In this report we have applied the Heisenberg picture to a set of problems including those discussed in Reference 4.

A. Free Space Propagation

The simplest example of the method is its application to the propagation of optical beams in free space, where the potential function V is a constant. The Heisenberg equations of motion for the transverse vector $\underline{\rho}$ become:

$$\frac{d}{dz}\underline{\rho} = \frac{1}{i\mathcal{K}}\left[\underline{\rho}, \frac{p^2}{2}\right] = \underline{p}$$

$$\frac{d\underline{p}}{dz} = \frac{1}{i\mathcal{K}}[\underline{p}, V] = 0 \tag{5}$$

The corresponding mechanical equations of motion are the same as the classical equations of motion for a particle in the absence of any applied force. The operator solution is (taking $z_0 = 0$),

$$\underline{\rho}(z) = \underline{p}(0) + z\,\underline{p}(0)$$
$$\underline{p}(z) = \underline{p}(0) \tag{6}$$

Thus, for the weighted average of the operator function $F(x, y, \partial_x, \partial_y; z)$ we have, equating the Schrodinger picture expression to the Heisenberg picture result,

$$\iint dxdy\, \mathcal{E}^*(x, y; z)\, F(x, y, \partial_x, \partial_y; z)\, \mathcal{E}(x, y; z) = \iint dxdy\, \mathcal{E}^*(x, y; 0)$$
$$F(x - i\mathcal{K}z\partial_x,\ y - i\lambda z\partial_y, \partial_x, \partial_y; z)\, \mathcal{E}(x, y; 0) \tag{7}$$

For the specific case where we require the moment $r_{mn}(0, z)$, we have

$$\iint dxdy\, x^m y^n |\mathcal{E}(x, y; z)|^2 = \iint dxdy\, \mathcal{E}^*(x, y; 0)[x - i\mathcal{K}z\partial_x]^m [y - i\mathcal{K}z\partial_y]^n \mathcal{E}(x, y; 0) \tag{8}$$

which is a polynomial in z of degree m+n. The lower moments have simple physical interpretations, e.g., for m = n = 0 we have conservation of energy:

$$\iint dxdy\, |\mathcal{E}(x, y; z)|^2 = \iint dxdy\, |\mathcal{E}(x, y; 0)|^2 \tag{9}$$

The second centrifugal moment $\iint dxdy(x^2 + y^2)\, |\mathcal{E}(x, y, z)|^2$ is a quadratic function of the propagation distance z. The physical significance of the lower moments is discussed by Vlasov, et al.[4] and Papoulis.[5]

B. Perturbation Theory -- Application to Propagation Through Random Media

The Heisenberg picture is particularly suited to the concise representation of the results of perturbation theory. Consider a wave equation of the form:

$$i\hbar \frac{\partial}{\partial z} \mathcal{S} = (H_0 + H_I)\mathcal{S} \tag{10}$$

where H_0 and H_I represent self adjoint operators. We wish to express averaged properties of the field developing under the full Hamiltonian in terms of expressions involving field averages under the unperturbed Hamiltonian, H_0. The results of perturbation theory are summarized by the following formal expression:

$$< F>_H = < (\exp \frac{+i}{\hbar} \int_0^z H_I(z')dz')_- \, F(z)(\exp \frac{-i}{\hbar} \int_0^z H_I(z')dz')_+ >_{H_0} \tag{11}$$

The subscript H on the field average on the left side of Eq. (11) indicates that the Heisenberg operator function $F(x(z), y(z), p_x(z), p_y(z); z)$ evolves in z under the full Hamiltonian $H_0 + H_I$, e.g., F satisfies the equation

$$\frac{d}{dz} F = \frac{1}{i\hbar} \left[F, H_0 + H_I\right] + \frac{\partial F}{\partial z} \quad \Big\} \text{ evolution under H} \tag{12}$$

The subscript H_0 on the field average on the right side of Eq. (11) indicates that the operators H_I and F evolve in z under just the unperturbed Hamiltonian H_0:

$$\left.\begin{aligned} \frac{dF}{dz} &= \frac{\partial F}{\partial z} + \frac{1}{i\hbar} \left[F, H_0\right] \\ \frac{dH_I}{dz} &= \frac{\partial H_I}{\partial z} + \frac{1}{i\hbar} \left[H_I, H_0\right] \end{aligned}\right\} \text{ evolution under } H_0 \tag{13}$$

The subscripts + and - on the exponentials in Eq. (13) are the time ordering symbols which indicate that in the power series expansion the non-commuting H_I factors are to be ordered in time positively (later to the left) and negatively (later to the right) respectively .

We now apply the general perturbation theory result, Eq. (11), to the problem of light propagation in a medium with random refractive index inhomogeneities. We assume the potential $V(\underline{\rho}, z)$ is a real random function. The moments of the mutual coherence function may be conveniently expressed in terms of the generating function $G(\underline{K}; \underline{\rho}_0; z)$, where

$$G(\underline{K}; \underline{\rho}_0; z) = < z | e^{i \underline{\rho} \cdot \underline{K}} \, e^{\frac{i}{\hbar} \underline{\rho}_0 \cdot \underline{p}} | z> \tag{14}$$

The formal perturbation expansion of this expression, treating the term $p^2/2$ as the unperturbed Hamiltonian H_0, is, in the Heisenberg picture:

$$G(\underline{K}, \underline{\rho}_0; z) = \frac{\overline{< 0 | (\exp\left[\frac{i}{\lambda}\int_0^z dz' V(\underline{\rho}(z'); z')\right])_- (\exp\left[\frac{-i}{\lambda}\int_0^z dz' V(\underline{\rho}(z') + \underline{\rho}_0 + (z'-z)\lambda \underline{K}; z')\right])_+}}{\overline{e^{i\,\underline{\rho}(z)\cdot\underline{K}}\, e^{\frac{i}{\lambda}\underline{\rho}_0 \cdot P}\, |0>}} \tag{15}$$

Equation (15) is a formally exact result, but approximation methods must be used in order to obtain practical expressions for G. The direct expansion of the time ordered exponential operators in powers of V is probably not as useful as a procedure which is capable of preserving the exponential form. As a direct generalization of the cumulant expansion of time-ordered exponentials[6] we note the following expansion to second order in K:

$$\begin{aligned} &\overline{\ell n(\exp \kappa \int_0^z \Omega_1(z')dz')_- (\exp -\kappa \int_0^z \Omega_2(z')dz')_+} \\ &= \kappa \int_0^z dz_1 \overline{(\Omega_1(z_1) - \Omega_2(z_1))} + \tfrac{1}{2}\kappa^2 \int_0^z dz_1 \int_0^z dz_2 \, \overline{(\Omega_1(z_1)\Omega_1(z_2))} \\ &- \overline{2(\Omega_1(z_1)\Omega_2(z_2))} + \overline{(\Omega_2(z_1)\Omega_2(z_2))}_+ - \overline{(\Omega_1(z_1) - \Omega_2(z_1))\,(\Omega_1(z_2) - \Omega_2(z_2))} \\ &+ \cdots\cdots \end{aligned} \tag{16}$$

We apply this approximation to Eq. (15) with the substitutions

$$\begin{aligned} \kappa\Omega_1(z') &= \frac{i}{\lambda} V(\underline{\rho}(z'); z') \\ \kappa\Omega_2(z') &= \frac{i}{\lambda} V(\underline{\rho}(z') + \underline{\rho}_0 + (z'-z)\lambda\underline{K}; z') \end{aligned} \tag{17}$$

At this stage we have to be specific about the statistical properties of the medium. We assume that $V(\underline{\rho}, t)$ is a random function with zero mean,

$$\overline{V(\underline{\rho}, z)} = 0 \tag{18}$$

and with a two point correlation function of the form:

$$\overline{V(\underline{\rho}, z)\, V(\underline{\rho}', z')} = A(\underline{\rho} - \underline{\rho}')\, \delta(z - z') \tag{19}$$

that is, a homogeneous process with zero correlation length in the propagation direction. Under this assumption the generating function for the moments becomes,

$$G(\underline{K}, \underline{\rho}_0, z) = \exp\left[-\frac{1}{\lambda^2}\int_0^z dz_1 [A(0) - A(\underline{\rho}_0 + \lambda K(z_1 - z)]\right] < 0 | e^{i\underline{K}\cdot\underline{\rho}} e^{\frac{i\underline{\rho}\cdot\rho_0}{\lambda}} | o> \quad (20)$$

and we have:

$$r_{oo}(\underline{\rho}_0, z) = G\Big]_{K=o} = r_{oo}(\underline{\rho}_o o) \exp\left[-\frac{A(0) - A(\underline{\rho}_0)}{\lambda^2}\right] z \quad (21)$$

$$r_{ol}(\underline{\rho}_0, z) = \frac{1}{i}\frac{\partial}{\partial Ky} G\Big]_{\underline{K}=o} = \left[r_{oo}(\underline{\rho}, o) - z\lambda \frac{\partial}{\partial y_o} r_{oo} + \frac{z^2}{2\lambda}\frac{\partial A}{\partial y_o} r_{oo}(\underline{\rho}_o, o)\right]$$

$$\exp\left(\left[-\frac{A(0) - A(\underline{\rho}_0)}{\lambda^2}\right] z\right) \quad (22)$$

etc.

Vlasov, et al.[4] obtained expressions for the moments of the mutual coherence function from a recurrence equation which was derived from an assumed wave equation for the mutual coherence function. The approximations which go into deriving the latter equation are parallel in our work, but in expressions that give the evolution of the moments directly.

Joint Services Technical Advisory Committee
F44620-74-C-0056

M. C. Newstein and D. Ramakrishnan

REFERENCES

1. For convenience of discussion we will restrict our consideration in this section to a system with 1 degree of freedom.

2. For a more complete discussion see a test in Quantum Mechanics such as "Quantum Mechanics, Vol. I: Fundamentals," by Kurt Gottfried (New York: W.A. Benjamin Inc., 1966).

3. The time dependence of the Schrodinger operator $F(x, p_x; t)$ is explicit, i.e., x and p_x are time independent operators.

4. S.N. Vlasov, V.A. Petrishchev, and V.I. Talanov, "Averaged Description of Wave Beams in Linear and Nonlinear Media (The Method of Moments)," Radiofizika, 14, a, 1353 (1971). This publication stimulated our present investigation when we saw how their free space results followed directly from the Heisenberg picture, and we attempted to use the same point of view on more difficult propagation problems.

5. A. Papoulis, "Ambiguity Function in Fourier Optics," JOSA, 64, 779 (1974).

6. R. Kubo, "A Stochastic Theory of Line-Shape and Relaxation," Scottish Universities Summer (1961), Ed. D. Ter Haar (New York: Plenum Press).

7. Operator methods are used in the discussion of propagation through random media, e.g., K. Furutsu, "Statistical Theory of Wave Propagation in a Random Medium and the Irradiance Distribution Function," June Opt. Soc. Am., 62, 240 (1972). The Heisenberg picture appears in this work but not in the simple content of the calculation of transverse averages as we have used it here.

TWO-PHOTON CORRELATIONS IN A WIDE ANGLE CHAOTIC LIGHT BEAM

S.M. Turner and D.B. Scarl

The difference in time of arrival between single photon events at two photomultipliers was measured in a quasimonochromatic spatially coherent widely divergent light beam.[1] A chaotic (Gaussian) light source was provided by a pure Ne^{20} discharge tube filled to a pressure of 1.6 torr and operated at a current of 15 mA. This experiment observed the Hanbury Brown-Twiss effect in light from the 585.2 nm line of neon 20, a line in which the effect had never before been seen, and measured the two-photon correlation function at much larger coherence angles than had previously been used.[2-10]

The coherence volume of an optical beam has usually been defined as the volume of a right cylinder whose base is the spatial coherence area and whose height is the temporal coherence length.[11] An angular coherence volume can be defined as the product of the solid angle subtended by the spatial coherence area (as viewed from the source) times the temporal coherence length. The angular coherence volume has the advantage of being a property of the source alone, since it is independent of the distance from the source to the downstream plane at which the spatial coherence is evaluated. It is also closely related to the volume of a single cell in phase space for the light leaving the source. For a circular aperture of diameter a, the coherence angle can be defined as λ/a.

Previous experiments have been carried out at small coherence angles and have therefore observed fields over which the range of $\underline{k}$ values was small, i.e., $|\Delta\underline{k}| < |\underline{k}|$. The largest coherence angle to date was 4 milliradians as compared with a coherence angle of 292 mr in the present experiment. This increased coherence angle allows the sampling of a much larger portion of the field leaving the source aperture by greatly increasing the range of $\underline{k}$ values observed. By an increase in the coherence angle our experimental purpose became threefold:

(1) As a check on the correctness of the chaotic density matrix to second order

(2) If we consider our coherence volume as corresponding to one cell of photon phase space, this experiment may provide additional information into the spatial and temporal extent of a photon

(3) As a first step in an experiment which will enable us to look at a small number of excited atoms.

Sidelight leaving the discharge tube was limited by a first circular aperture that consisted of a 2 micrometer diameter hole in a 1 μm thick nickel foil. A second aperture, with a diameter of 12.7 mm placed 50 mm downstream provided the effective detector size and allowed the photomultipliers to detect substantially spatially coherent light. A 585.3 nanometer interference filter with peak transmission of 50% and a full width at

half maximum of 3.0 nm assured that only the 585.2 nm line was observed. This Doppler broadened line was chosen for its high spectral radiance; the results of a scanning Fabry-Perot interferometer gave a spectral width of 2108 ± 50 MHz (FWHM).

By a (delayed) coincidence technique with a dual discrimination system and two photomultipliers having a Gaussian time resolution function with a full width at half maximum of 1.5 ns, a time to height converter recorded a coincidence count whenever two photon events occurred at both detectors within a time interval of 20 ns of each other. The output pulses from the time to height converter were analyzed and stored in a pulse height analyzer calibrated to 100 ps per channel according to the difference in the time of arrival of the single photon events.

The experimental results are shown in Fig. 1, together with a curve of the two-

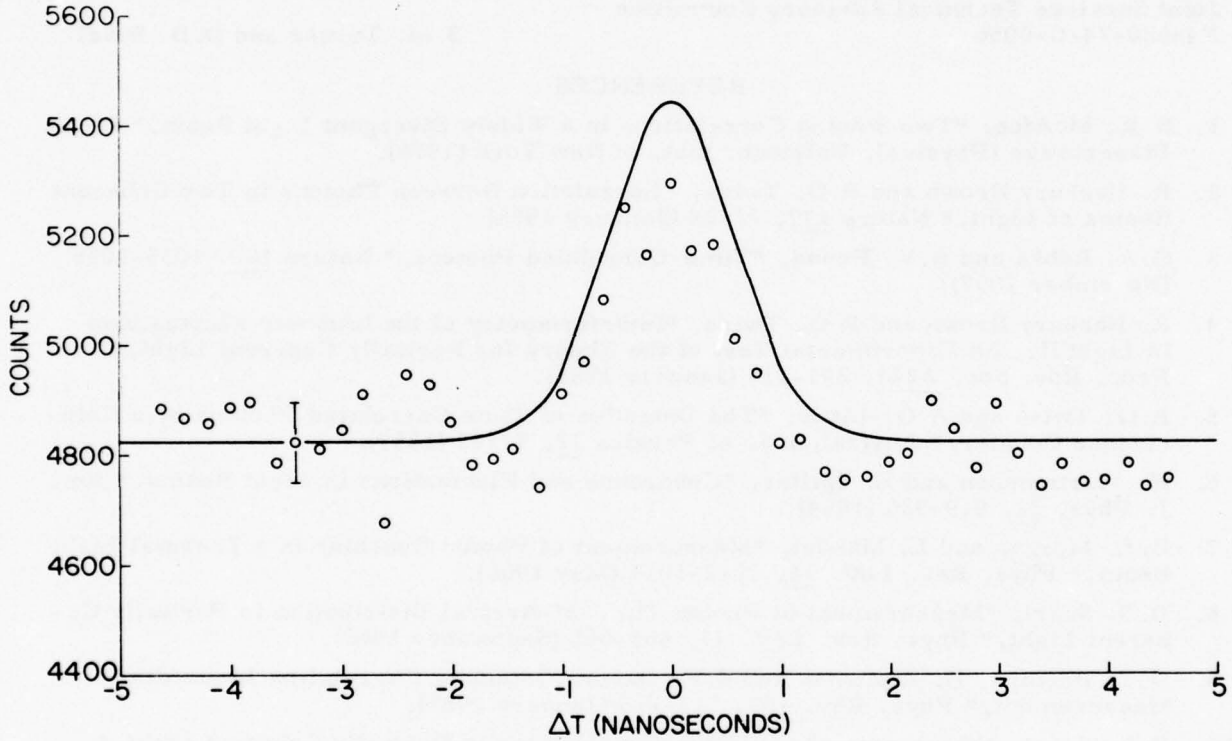


Fig. 1. The two-photon counting rate $R(\Delta t)$ vs. Δt for light from a neon discharge passing through a 2 micrometer diameter aperture. The solid line is the prediction of a maximum entropy density matrix. The error bar shown is common to all of the experimental points.

photon correlations to be expected on the basis of a constant energy maximum entropy density matrix for the chaotic field. The measured points and calculated curve are in

reasonable agreement, showing that the usual chaotic density matrix provides a good representation of the chaotic electromagnetic field for coherence angles at least as large as 292 mr.

The Hanbury Brown-Twiss effect was observed in the 585.2 nm spectral line of neon for a coherence angle of 292 mr, representing a solid angle approximately 4000 times greater than that used in any previous second order correlation function measurement. The large coherence angle (1) served as a check on the choice of the maximum entropy constant energy density matrix to second order by allowing the observation of a much larger portion of the field leading to an increase in the region of the coherence volume, and (2) allowed the investigation of the effect on the counting statistics from small source apertures, i.e., large coherence angles.

Joint Services Technical Advisory Committee
F44620-74-C-0056

S.M. Turner and D.B. Scarl

REFERENCES

1. S.R. McAfee, "Two-Photon Correlations in a Widely Divergent Light Beam," Ph.D. Dissertation (Physics), Polytech. Inst. of New York (1976).
2. R. Hanbury Brown and R.Q. Twiss, "Correlation Between Photons in Two Coherent Beams of Light," Nature 177, 27-29 (January 1956).
3. G.A. Rebka and R.V. Pound, "Time-Correlated Photons," Nature 180, 1035-1036 (November 1957).
4. R. Hanbury Brown and R.Q. Twiss, "Interferometry of the Intensity Fluctuations in Light II. An Experimental Test of the Theory for Partially Coherent Light," Proc. Roy. Soc. A243, 291-319 (January 1958).
5. R.Q. Twiss and A.G. Little, "The Detection of Time Correlated Photons by a Coincidence Counter," Australian J. of Physics 12, 77-93 (1959).
6. W. Martienssen and E. Spiller, "Coherence and Fluctuations in Light Beams," Am. J. Phys. 32, 919-926 (1964).
7. B.L. Morgan and L. Mandel, "Measurement of Photon Bunching in a Thermal Light Beam," Phys. Rev. Lett. 16, 1012-1015 (May 1966).
8. D.B. Scarl, "Measurement of Photon Time-of-Arrival Distribution in Partially Coherent Light," Phys. Rev. Lett. 17, 663-666 (September 1966).
9. D.T. Phillips, H. Kleinman and S.P. Davis, "Intensity Correlation Linewidth Measurement," Phys. Rev. 153, 113-115 (January 1967).
10. D.B. Scarl, "Measurement of Photon Correlations in Partially Coherent Light," Phys. Rev. 175, 1661-1668 (November 1968).
11. L. Mandel and E. Wolf, "Coherence Properties of Optical Fields," Rev. Mod. Phys. 37, 231-287 (April 1965).

AVALANCHE INITIATED OSCILLATIONS IN UNIPOLAR STRUCTURES UNDER PULSED EXCITATION

B. Senitzky and S. Gottfried

Avalanche in semiconductor devices became an interesting phenomenon to many people when J.B. Gunn[1] showed that the avalanche region in a semiconductor could, under the proper conditions, cause a negative differential resistance in the devices' terminal characteristics. Useful sources of microwave energy up to millimeter frequencies were developed,[2-4] based on this phenomenon. These devices were called IMPATT diodes (an acronym for Impact Avalanche Transit Time).

Several years later, a negative resistance device was discovered which involved a different mechanism, but was avalanche initiated. Dubbed the TRAPATT (Trapped Plasma Triggered Transit), it was shown to oscillate at lower frequencies than the IMPATT, but at much higher efficiency.[5]

At around the same time, some interest was generated in studying avalanche phenomena in unipolar devices. Several authors[6] had predicted the existence of a negative differential resistance in near intrinsic silicon and germanium. In 1972, Dworsky and Harrison[7] proposed an analytical model for an n^+nn^+ structure under avalanche conditions. The analysis provided an explanation for the differential negative resistance but there was little experimental evidence to support the existence of the phenomenon.

This contribution is a report of the experimental, analytical and computer investigation to observe and study the avalanche initiated phenomena occurring in an n^+nn^+ structure under pulsed conditions. The motivation for the work is twofold: the n^+nn^+ structure is more simple to fabricate than a junction device, and therefore has the potential of being a more attractive, useful source of microwave energy, and second the study of the n^+nn^+ device could improve our understanding of ohmic contacts under high fields, a situation that occurs very often in solid-state device and integrated circuit technology.

A. Device Operation

This is a simplified explanation of the physics of the phenomena that occur in an n^+nn^+ structure when excited with a high-speed, high-current pulse. The effects following the excitation can be divided into three periods:

(1) A charging period where the voltage across the device increases rapidly

(2) A plasma formation period in which the diode switches to a low voltage, high-conductivity state

(3) An extraction period, in which the plasma is swept from the active region and the diode voltage and resistivity increase.

B. Changing Period

Consider an n^+nn^+ structure, in which a current flows which requires a larger electron concentration than exists in the n region. The electric field will then assume a triangular profile, as shown in Figure 1. Consider a high-speed pulse of current, with amplitude large enough to cause avalanche. (The speed and amplitude of the current pulse are defined in terms of the semiconductor properties and the device geometry.) In order to support this high-current level, a pulse of electrons will be injected from the cathode into the n region. Ahead of this pulse, the current is supported by a uniform rise of the electric field (i.e., a displacement current).

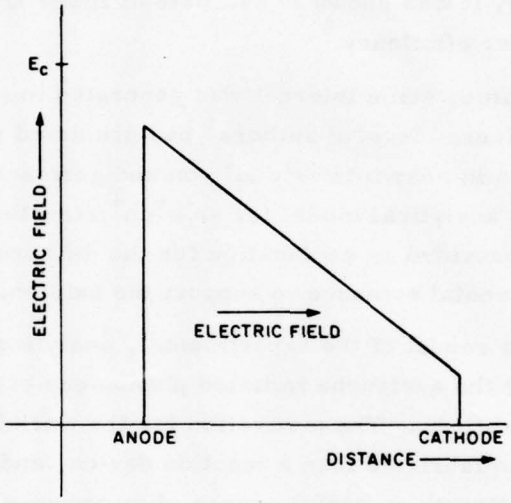


Fig. 1. Electric field profile of n^+nn^+ structure under high current conditions.

Since the field peaks at the anode (Fig. 1), the rapid increase will cause avalanching to occur there first (assuming the rise is rapid enough so that the electron pulse from the cathode has not reached the anode yet). As the field keeps increasing, the point at which avalanche occurs will sweep across the n region (from anode to cathode) at a velocity which is different than the saturated carrier velocity. (For high efficiency this velocity should be greater.)

C. Plasma Formation Period

At the avalanche point, a very high density of electrons and holes form. For the above conditions, the holes are left behind the avalanche zone. Since the electrons will drift towards the anode and holes will follow behind the avalanche zone towards the

cathode, the result is a net positive charge behind the moving zone. This causes the slope of the field to reverse and decrease behind the moving zone. For the device parameters considered here the field at this point will drop to a small value.

Under the influence of the small electric field behind the zone, a high density, slowly moving plasma is formed and spreads out over most of the device. When the moving avalanche zone meets the electron pulse from the cathode, the avalanche will quench and there is no generation of charges. The result is a low voltage and high current in the device terminal characteristic. Since the voltage has decreased while the current remained constant, a differential negative resistance is realized.

D. Plasma Extraction Period

Under the influence of the small electric fields, the electrons and holes will move slowly towards the anode and cathode, respectively. In the region behind the moving holes, the net space charge is negative and the electric field will assume an increasing negative slope. As the slope increases, the velocity of the carriers will increase and the terminal voltage is rising. Since the current pulse is held constant, the result is an increase in device terminal resistance.

The device terminal characteristics calculated by a piecewise linear analysis based on the above arguments are summarized in Figure 2. It can be shown that the ratio of

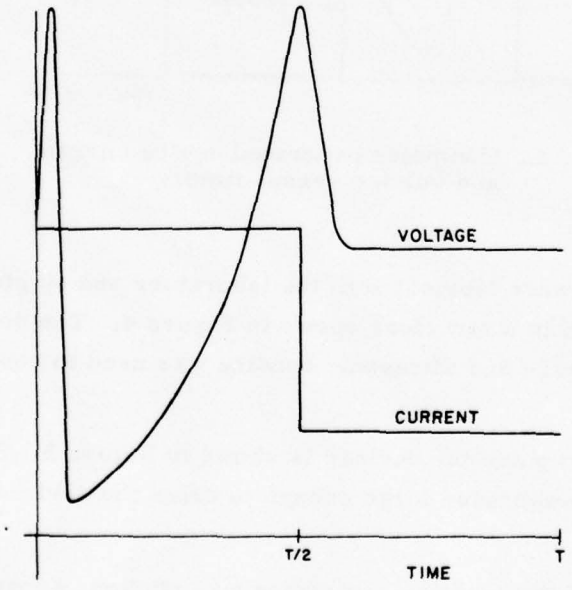


Fig. 2. Theoretical device current and voltage versus time.

the first harmonic components of the current and voltage are out of phase, and hence the device will exhibit a differential negative resistance in its terminal characteristic.

A numerical solution on a computer of the Maxwell Equations in one dimension for the boundary conditions of an n^+nn^+ structure resulted in the waveforms shown in Figure 3. The computer generated waveforms are similar to the ones calculated in the piecewise linear approximation summarized in Figure 2. The numerical analysis is not limited by many of the assumptions of the piecewise linear theory and agreement provides good support of the simplified analysis.

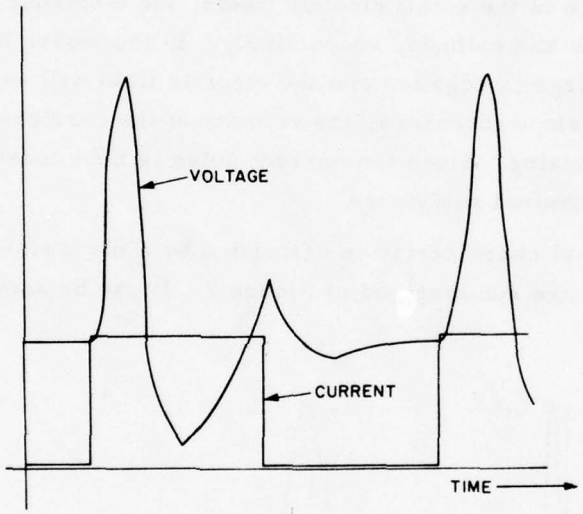


Fig. 3. Computer generated device current and voltage versus time.

E. Experiment

The n^+nn^+ wafers were fabricated in the laboratory and single devices cut from the wafers were mounted in a structure shown in Figure 4. The devices were bonded to the bottom of the package and ultrasonic bonding was used to contact the top of the device.

The circuit used to pulse the devices is shown in Figure 5. Subnanosecond rise-time current pulses of amplitudes large enough to drive the devices into avalanche are available.

The mechanism of the negative resistance was studied, as well as the effects of device thickness, area, surface preparation, current amplitude, and doping density on

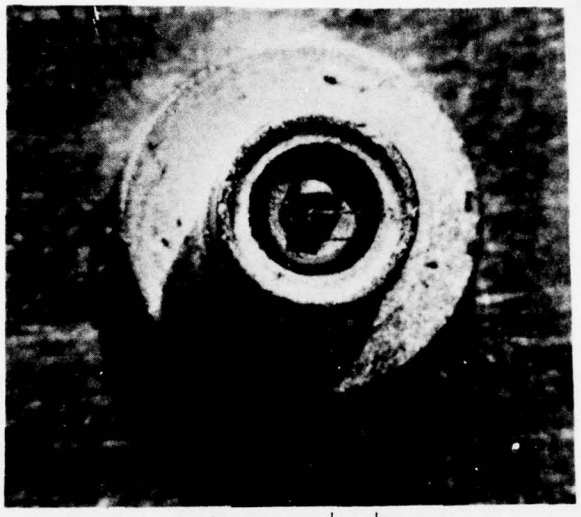


Fig. 4. Packaged n^+nn^+ device.

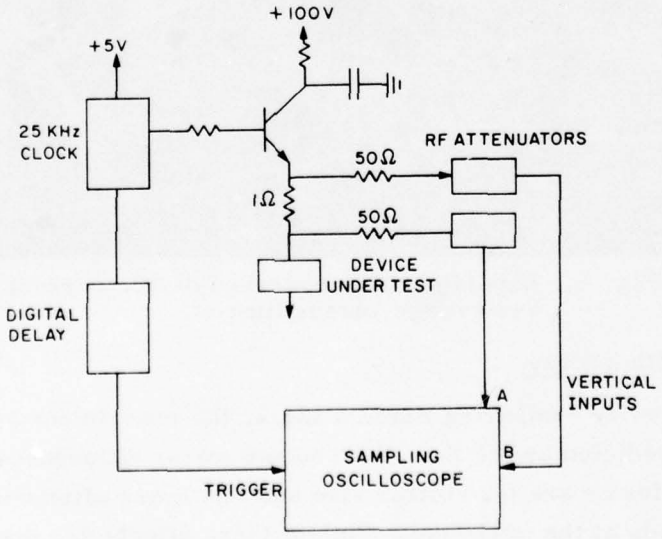


Fig. 5. Pulsing circuit for devices.

the rise time, plasma extraction time, negative resistance, efficiency, lifetime, and stability of the devices fabricated.

F. Results

The pulsed I-V characteristics as observed using a sampling oscilloscope of the trapped plasma devices under pulsed conditions is shown in Figure 6. This is the first time that these results have been obtained. The photograph shows the pulsed current excitation and the resultant terminal voltage as a function of time.

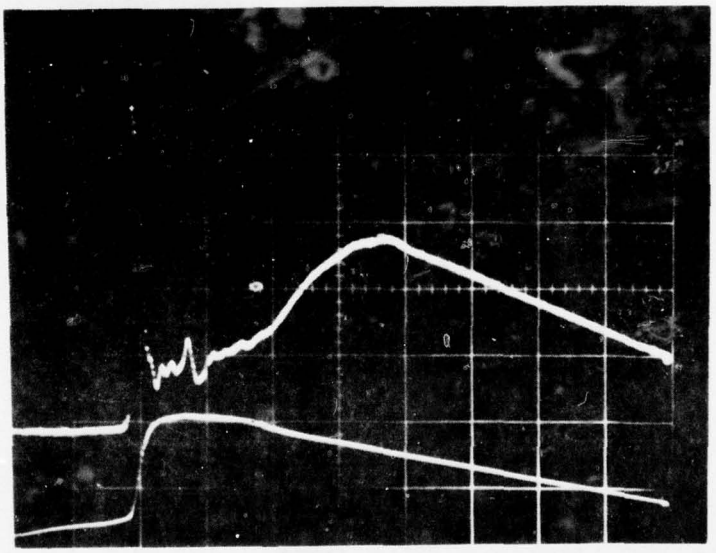


Fig. 6. Experimentally observed device current and voltage versus time.

G. Comparison with Theory

As can be seen by comparing Figs. 6 and 2, the experimental waveforms are similar to those predicted by the theory. The two major differences between the theory and observed waveforms are the voltage rise and fall times after current pulsing and the voltage amplitude at the minimum. Both of these effects are due to the 0.4 nanosecond rise time of the sampling oscilloscope used. At the present time, the voltage

waveforms are being deconvolved with the impulse response of the oscilloscope, to determine the original voltage signals. The results of this operation should show a closer agreement between the theory and experiment.

The harmonics of the observed waveforms are presently being analyzed to determine the magnitudes of the negative resistance observed and a predicted oscillation efficiency (as determined by the relative magnitude of the first harmonic component).

H. Conclusion

The experimental results show good agreement with the theoretically predicted terminal voltages and currents of the n^+nn^+ device under pulsed excitation. The device is shown to be a useful source of microwave energy.

Joint Services Technical Advisory Committee
F44620-74-C-0056

B. Senitzky and S. Gottfried

REFERENCES

1. J.B. Gunn, "Avalanche Injection in Semiconductors," Proc. Physical Society (London), B69, 781-790 (1956).
2. W.T. Read, Jr., "A Proposed High Frequency Negative Resistance Diode," Bell System Technical Journal, 37, 401 (1957).
3. R.L. Johnston, B.C. DeLoach and B.G. Cohen, "A Silicon Diode Microwave Oscillator," Bell Systems Techical Journal, 44, 369 (1965).
4. C.A. Lee, R.L. Batdorf, W. Wiegmann and G. Kaminsky, "Technological Developments Evolving from Research on Read Diodes," IEEE Trans. on Electron Devices, ED-13, No. 1, 175 (January 1966).
5. H.J. Prager, K.K.N. Chang and S. Weisbrod, "High Power High Efficiency Silicon Avalanche Diodes at Ultra High Frequencies," Proc. IEEE, 55, 586-687 (April 1967).
6. J.N. Park, K. Rose and K.E. Mortenson, "Avalanche Breakdown in Near Intrinsic Silicon and Germanium," J. Appl. Phys., 38, No. 13 (1967).
7. L.N. Dworsky and R.I. Harrison, "Trapped Plasma Oscillations in Unipolar Semiconductor Structures," IEEE Trans. on Electron Devices, ED-19, No. 6, 836-838 (June 1972).

HALL EFFECT IN SILICON INVERSION LAYERS

Y. L. Yao and B. Senitzky

The effect of transverse magnetic field on a silicon inversion layer has been studied[1] at liquid nitrogen temperature. The results, similar to that of the Hall effect, showed that the channel conductance varied both linearly and quadratically with the applied magnetic field. Using thick oxide MOS devices, the maximum change of channel conductance of 10 percent has been observed at 15 kG. This observed value is much larger than the theoretical values obtained from either the classical model[2] or quantum mechanical analysis.[3]

This phenomenon could be of interest in the study of charge coupled devices (CCD).[4] It has been established that one of the factors which prevent ideal charge transfer in a CCD is the loss of charge to surface states at the Si-SiO_2 interface. The variation of channel conductance with magnetic field suggests that the moving carriers in the inversion layer are being pushed toward or away from the surface by the Lorentz force, thus causing a decrease or increase of conductance. However, devices used in previous experiments[1] were not suitable for DDC's because of the thick gate oxide. Consequently, we experimented with MOS devices with gate oxide's whose thickness comparable to that of CCD's.

In this contribution we will describe our measurements of channel conductance variation of a MOS device at liquid nitrogen temperature and magnetic fields up to a maximum of 15 kG. We will then compare these measurements to surface scattering model and show that this model is in good agreement with our experimental data.

A. Previous Physical Interpretations

A first order classical analysis of this Hall effect was first given by Tansal.[2] He stated that the number of carriers in an inversion layer was modified by the magnetic field and thus resulted in a change of channel conductance. However, there are several problems with this approach. First, it is not clear how the number of carriers in an inversion layer is affected by magnetic field. Secondly, this approach results only in a linear relation of the channel conductance and the magnetic field.

A quantum mechanical approach to determine the Hall voltage of a thin current layer under a magnetic field was given by Stern.[3] He limited his analysis to the case[5] where all carriers are in the lowest level ($n = 0$). The average distance of the carriers from the surface is calculated both for zero and finite magnetic field. Using second order perturbation theory, the change of this distance with magnetic field is calculated. The modified surface potential can then be computed from this change. This modified potential is referred by Stern as the quantum mechanical Hall voltage.

A first order estimate of the channel conductance variation can be made by adding the quantum mechanical Hall voltage directly to the effective gate voltage and computing the conductance from the relation between the conductance and the gate voltage. However, the computed conductance change is much smaller than the observed value.

B. Surface Scattering Model

We propose a different model which accounts for the changed carrier mobility. It is well known that carriers in an inversion layer are diffusely scattered from the oxide surface and thus show lower mobility than bulk material. We used a simple classical model similar to that suggested by Fang and Triebwasser[6] and assumed that:

(1) the density of the free carriers is uniform in the channel, and

(2) the gate oxide is free of charge.

To simplify the calculation, we selected a carrier initially at rest at the interface of the inversion and the depletion layers and moving toward the surface. We then calculate the time for the carrier to reach the surface with and without a magnetic field.

Assuming that this carrier reaches the surface with no intermediate scattering and its effective mass remains constant, we can compute the mobility from this transit time. The resultant channel conductance variation due to magnetic field shows both the linear and quadratic relationship to the magnetic field. More important, the general shape of the resultant curve agrees well with observations as will be shown.

C. Experiments and Results

Figure 1 shows the device structure used in this study. The structure is a linear n-channel MOS transistor with source-drain spacing of 2 mil and a channel width of 40 mil. The channel is measured at the source contact. The rectification properties of the source and drain diffusions allow only one type of carriers (the majority carriers or electrons in this case) to flow. Figure 2 shows the electrical circuit used in the experiment.

The devices were mounted on a non-magnetic substrate with brass pins for electric connections. A special liquid-nitrogen dewar was designed to allow the device to be placed in a Varian electromagnet with a 20 mm airgap. Figure 3 shows the physical experiment set-up.

A typical curve showing the variation of the dc source-drain current as a function of magnetic field at fixed source-drain and gate voltages is shown in Figure 4. The vertical bar at each data point represents the variation of each measurement due to noise. The smooth curve is a least square fit of the experimental data to a second order polynomial.

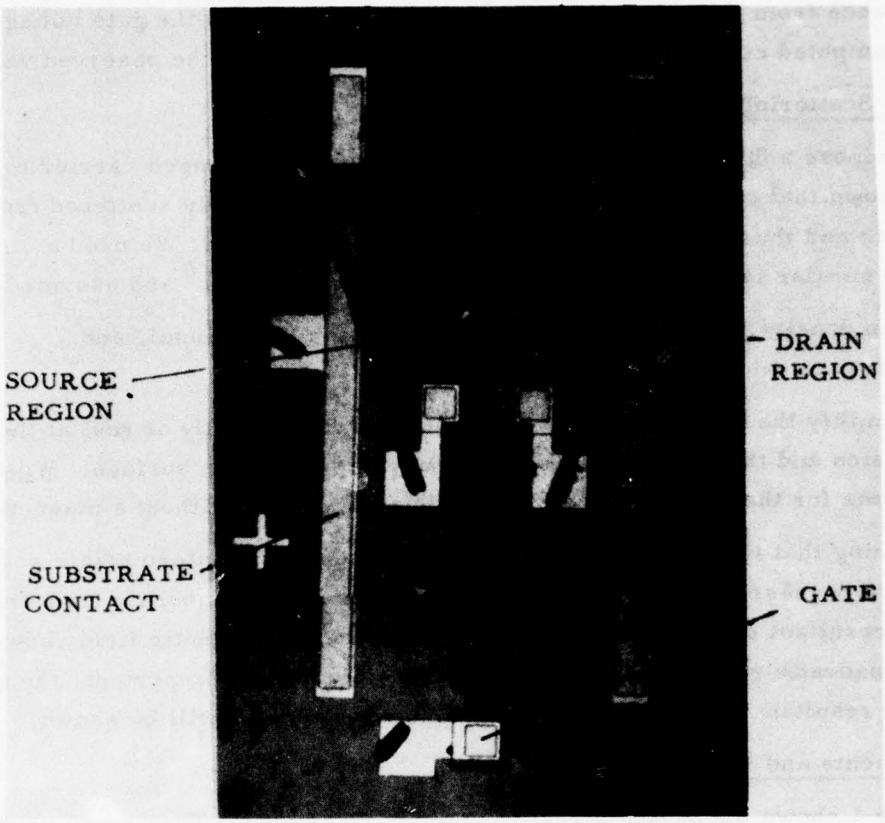


Fig. 1. MOS device structure.

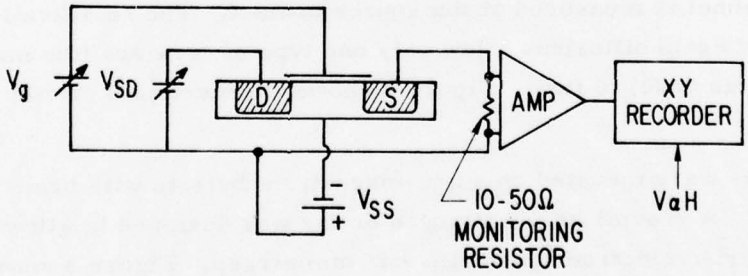


Fig. 2. Experiment set-up for dc channel conductance measurement.

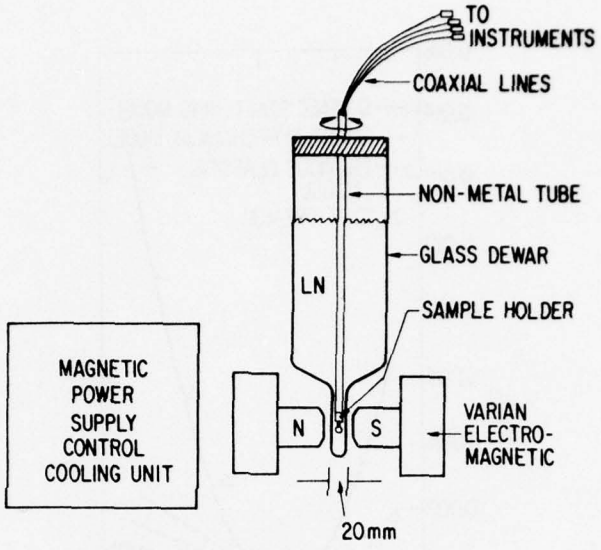


Fig. 3. Experiment set-up.

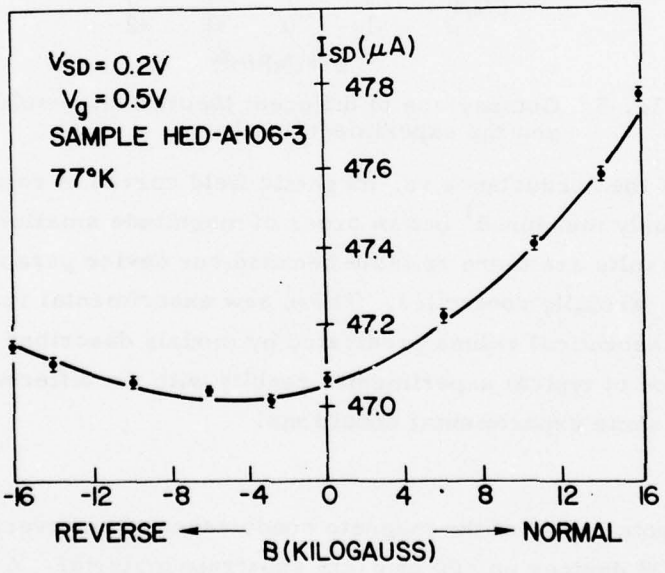


Fig. 4. Typical dc source-drain current variation vs. magnetic field at fixed source-drain and gate voltages.

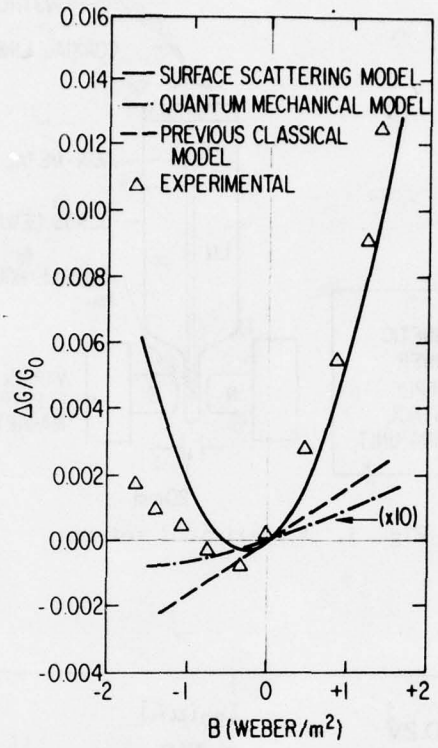


Fig. 5. Comparison of different theoretical results and the experimental values.

The shape of the conductance vs. magnetic field curves is reproducible and is similar to that previously measured[1] but an order of magnitude smaller. In our opinion we feel that our results are more reliable because our device parameters, such as oxide quality, are more carefully controlled. These new experimental results are in closer agreement to the theoretical values predicated by models described above. Figure 5 shows a comparison of typical experimental results with the different theoretical values calculated for the same experimental conditions.

D. Summary

An experimental study of the magneto conductance of an inversion layer was made with N-channel MOS devices on 100 ohm-cm substrate material. A 1.2 percent change in channel conductance was observed at a dc magnetic field of 15 kG applied tangentially to the surface and perpendicularly to the current flow. This is an order of magnitude smaller than previous experimental results[1] using similar devices with thick oxide gates.

A surface scattering model based on the theory of variable carrier mobility was presented. This model yields both the linear and quadratic dependence of channel conductance on magnetic field. The agreement with the experimental results is much better than that achieved with previous models. Nevertheless, we feel that further experiment is required in this area to determine the exact nature of this phenomena.

Joint Services Technical Advisory Committee
F44620-74-C-0056

Y. L. Yao and B. Senitzky

REFERENCES

1. S. Tansal, A.B. Fowler and R.F. Cotellessa, Phys. Rev., 178, 1326 (1969).
2. S. Tansal, Ph.D. Thesis, New York University (1969).
3. F. Stern, Phys. Rev. Letters, 21, 1687 (1968).
4. W.S. Boyle and G.E. Smith, "Charge Coupled Semiconductor Device," B.S.T.J., 49, 587-593 (1970).
5. F. Stern and W.E. Howard, "Properties of Semiconductor Surface Inversion Layers in the Electric Quantum Limit," Phys. Rev., 163, No. 3, 816-835 (1967).
6. F.F. Fang and S. Triebwasser, IBM J. of Res. and Dev., 8 (1964).

LARGE SIGNAL TRANSIENT RESPONSE OF MOSFETS

R. Kinasewitz and B. Senitzky

Although the metal-oxide-semiconductor field-effect transistor (MOSFET) has been in use for a long time, most studies to date have been directed towards its steady state and small signal operation. The reason for this is the difficulty of analyzing the nonlinear physical phenomena involved in the large signal transient conditions.[1]

All investigations formerly published on the transient behavior of the MOSFET have been subjected to various restrictions. For example, some investigators assume the drain-to-source voltage to be zero,[2] or consider only very long channels where the parasitic capacitances are of little influence.[2,3] Others use oversimplified physical assumptions.[4]

Our investigation is directed towards identifying and quantifying some of the physical phenomena which are particularly significant with respect to the turn-on behavior of MOSFETS.

A. Experiment

Our experiment consists of exciting a MOSFET with a step input in the circuit configuration shown in Fig. 1 and measuring the output. The excitation is obtained from the negative step calibration output of a Tektronix Type 519 Oscilloscope. This step waveform has a rise time of about 100-150 picoseconds. The MOSFET circuit of Fig. 1 is installed in a General Radio Type 874-X Insertion Unit. The output waveform is observed on a Taktronix Type 564 Storage Oscilloscope containing the Type 3T77 Sampling Sweep and Type 3576 Vertical Sampling plug-in units. The Type 3576 plug-in unit has a specified rise time of less than 400 picoseconds.

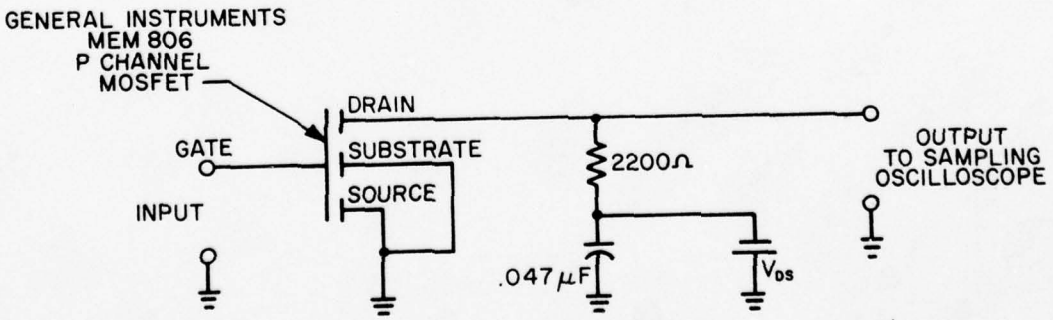
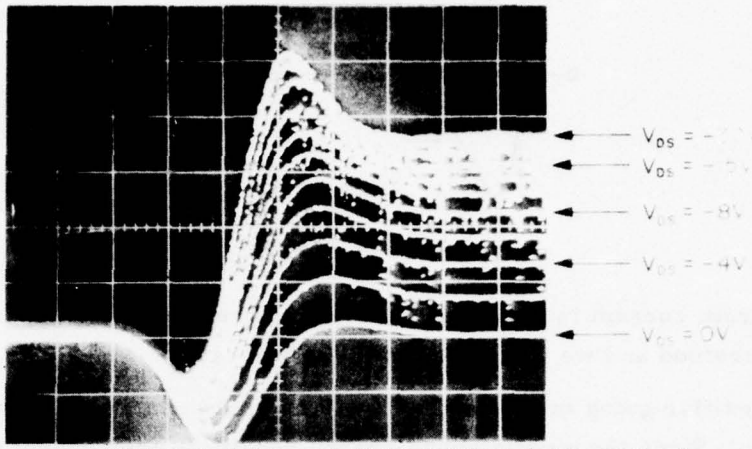


Fig. 1.

B. Results

A typical set of output waveforms is shown in Figure 2.



VERTICAL AXIS: VOLTAGE OUTPUT (500 mV/DIV)
HORIZONTAL AXIS: TIME (0.6 nsec/DIV)
INPUT SIGNAL: -8V STEP

Fig. 2.

The negative-going undershoot portion of the response has been observed and is thought to arise from two mechanisms.[2,5]

One mechanism is the parasitic capacitances: the gate-to-source capacitance (C_{GS}), the gate-to-drain capacitance (C_{GD}), and the drain-to-source capacitance (C_{DS}), (see Figure 3). The parasitic capacitance effect on the MOSFET transient can be estimated by considering the circuit of Fig. 4 and noting that the output response of this circuit to a negative step input is that shown in Fig. 5, which is similar to the observed undershoot.

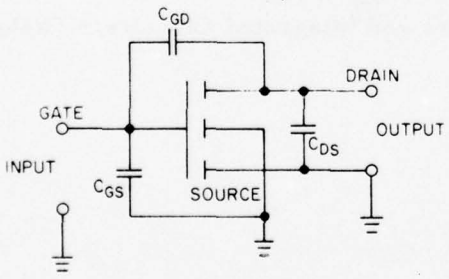


Fig. 3.

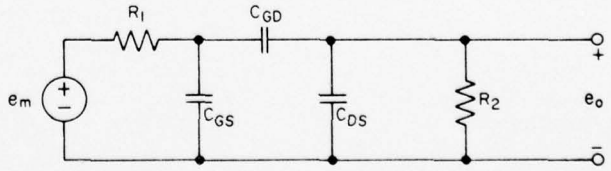


Fig. 4.

The other proposed reason for the undershoot is that at the beginning of the transient the holes in the inversion layer are partially supplied from the drain and therefore

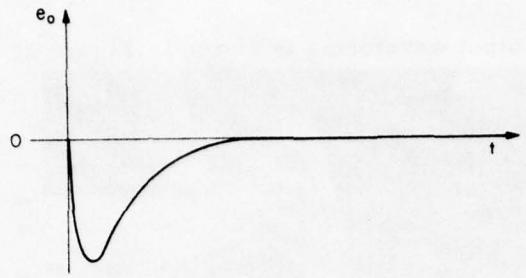


Fig. 5.

the initial drain current is reversed. This drain current reversal is not widely known or well understood and we would like to investigate it.

The positive going overshoot of Fig. 2 has not been observed and may be due to the following: Since the mobile space charge density in the channel inversion layer is a function of the drain-to-source voltage,[6] the channel has to acquire and dispose of part of its mobile space charge before reaching steady state.

Work will be continued on the above phenomena and to this end provisions are being made to use a faster detection system.

Joint Services Technical Advisory Committee
F44620-74-C-0056

R. Kinasewitz and B. Senitzky

REFERENCES

1. T.W. Collins, Ph.D. Dissertation, University of California, Davis (June 1973).
2. K. Goser, Arch. El. Ubertr., 24, 21 (1970).
3. D. Landgraf-Dietz, Nachrichtentechnik, 18, 296 (1968).
4. M.E. Zahn, "Solid-State Electronics," 17, 843 (1974).
5. F. Grimmer and K. Goser, Arch. El. Ubertr, 26, 197 (1972).
6. P. Richman, "MOS Field-Effect Transistors and Integrated Circuits," (Wiley-Interscience, 1973).

RELIABLE OHMIC CONTACTS TO SILICON DEVICES

M. Eschwei and S. Gottfried

An important step in the fabrication of silicon devices is the formation of an ohmic contact to the device. Silicon oxides[1-3] and other contaminants on the surface,[4,5] as well as processing induced impurities, will preclude reliable contacts with good electrical characteristics.

Many authors have studied the processes, materials, and results of making ohmic contacts to semiconductor structures.[2,5] Reliable adhesion of the metal to the silicon, and ease of bonding to the contact are of major concern. Gold is a desirable contact material because of its resistance to corrosion, ease of bonding, and high elongation (allowing thermal expansion mismatch with the silicon).[6] The adherence of gold, and all precious metals, to silicon is however, very poor. Chromium has been used as a precoating to improve the adhesion.[5] Sputtering several layers of material[6,7] provides a good contact to the silicon, but is a complicated technique. Glow discharge cleaning just prior to film deposition improves adhesion, but this may add interface contaminants.[8] Other processes, such as electroless nickel plating,[9] are less reliable.

Having tried various wafer cleaning and etching processes, and contact formation by thermal evaporation and electroplating of different metals, we have developed a simple, reliable technique which does not require any expensive doped gold targets. It provides ohmic contacts to silicon devices with excellent adhesion and repeatable electric properties.

We decided to try the sputter-etch technique of removing surface impurities for obtaining good ohmic contacts with strong adhesion,[10] followed immediately by hot filament evaporation of the contact material. Using this process we avoid buying an expensive, limited use doped gold target. A Materials Research Corporation 8620 RF Sputtering Module which was collar-mounted on a 6" oil diffusion pumped system was modified. A high current feed thru has been installed in the sputtering module collar to supply power to a resistively heated gold-antimony source. This filament source is mounted in a sheet stainless steel housing located next to the etching platform in the sputtering unit. Proper shielding keeps the rest of the system from being coated. The stranded tungsten filament is roughly "W" shaped to fit the small area available. The filament is wound with 18 cm of 0.25 mm diameter 99.95 gold wire over 76 cm of 0.10 mm diameter gold wire doped with 0.6 percent antimony. After detergent and solvent ultrasonic cleaning, it is mounted in the filament housing and the system is evacuated for premelting in vacuum. This removes any surface impurities which might be left after ultrasonic cleaning, and it also makes the gold-antimony wet and flow along the filament for better coverage of the wafer.

A "pusher" has been fastened to the movable shutter supplied with the module. Using this shutter, the wafer, as shown in Fig. 1, can then be pushed from the sputter-

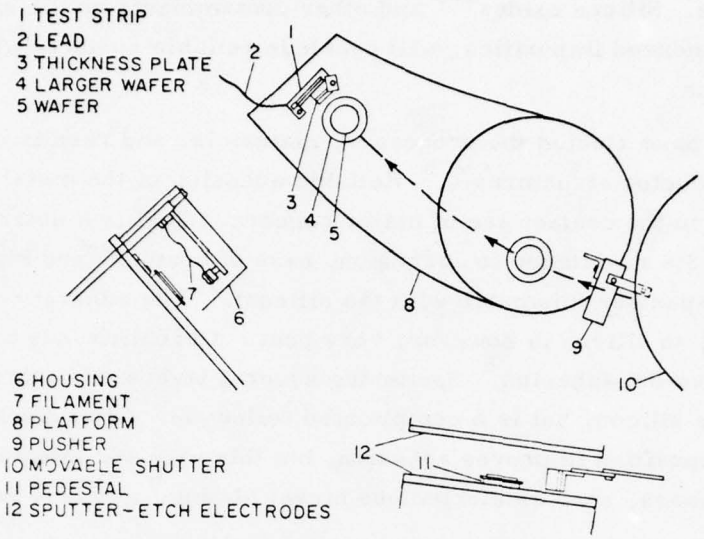


Fig. 1. Diagram of parts added and setup.

etch platform over and under the filament for downward film deposition immediately *after sputter-etching*. The wafer to be made into a device is itself placed on a larger silicon wafer which in turn is on a "pedestal" of 3 mm thick glass. The larger wafer protects the device wafer from any back-sputtering impurities from the "pedestal" or the aluminum target plate. The thick glass raises the wafers and provides enough thickness for the "pusher" to contact and move the wafer into position under the filament.

Mounted on a demountable platform under the filament housing is a glass test strip with fired-on platinum-gold contacts. A mica mask is used to give a film area of about 40 squares. This allows the film thickness to be monitored during the evaporation, since it results in a film with a resistance high enough to be read on an ohmmeter for a thickness of 6000 to 7000 angstroms. A partially masked piece of microscope slide glass is also mounted on the platform so that the film thickness step can be checked out later by both Tolansky interferometry[11] and Talysurf profilometry[12] measurements.

The glass test strip and the thickness plate are cleaned and mounted on the demountable platform, which is then put in the system. Different cleaning methods were used for the wafer, after which it was immediately placed on the "pedestal" and the

system evacuated. Standard procedures were used for pumping down and flushing the system with ultra pure argon. Approximately 5500 angstroms of the n-type silicon wafer was etched from the top surface at 150 Watts forward power applied for 55 minutes. An argon atmosphere is maintained at 0.86 Pascals (6.5 microns) by throttling back the high vacuum valve. Immediate opening of the high vacuum valve on completion of the etching brings the pressure down to less than 8.8×10^{-4} Pascals (6.6×10^{-6} Torr). The wafer is then pushed under the filament for the immediate deposition of the antimony-gold. Approximately 80 amps filament current brings the test strip resistance to about 4 Ohms in 5 seconds, corresponding to a film thickness of approximately 7000 angstroms.

The two sides of the wafer must be done in separate runs. There is no apparent degradation of the first side with the sputter-etch and deposition of the second side.

The sputter-etch plus vacuum deposition technique was used to form contacts on n-type epitaxial silicon wafers. The 50 Ohm centimeter resistivity epitaxial layer is formed on an n^+ substrate, with a resistivity of .001 Ohm centimeter. Two sets of wafers were prepared. One set was etched approximately 6.5 minutes in Buffer HF solution,[13] and the second set was etched 15 seconds in a solution of nitric, hydrofluoric, and acetic acids. The Buffer HF removes only surface oxides, while the acid mixture actively etches the wafer. The wafers were thoroughly washed in deionized water and blown dry with dry nitrogen. As described above, these wafers were then sputter-etched, followed by the vacuum deposition of the antimony doped gold wire. As a control, contacts were formed on two other sets of wafers using the same chemical-etch cleaning procedures, except that these were not sputter-etched. They were simply loaded into a conventional vacuum system, immediately evacuated, then the antimony doped gold film was deposited from a molybdenum strip boat by thermal evaporation at 8.8×10^{-4} Pascals.

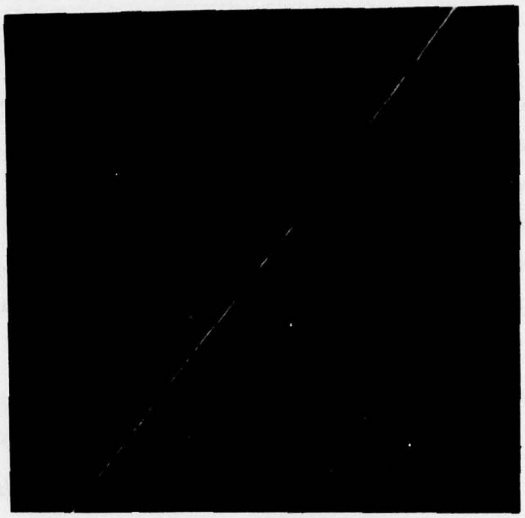
The I-V characteristics of the four sets of slices were compared using a diode curve tracer. The typical I-V characteristics of the wafers that were sputter-etched are shown in Figure 2a. The control groups' I-V characteristics were typically that as shown in Figure 2b. As can be seen, those with the sputter etching done prior to the vacuum deposition provide a good ohmic contact, while the others have very "leaky" blocking type contacts. The results were very consistent.

The adhesive tape tests[14] and scratch tests[15,16] for film adhesion were performed on all the wafers. The sputter etched groups were the only ones that maintained good adhesion through out all the tests. The gold contact on the other wafers peeled off with adhesive tape, and with diamond scoring, making it impossible to maintain contacts to make devices. Leads were easily attached to the strongly adherent gold contacts by thermocompression bonding.

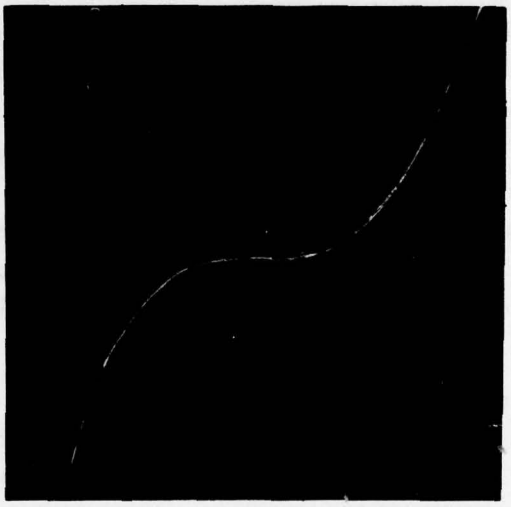
AD-A034 764 POLYTECHNIC INST OF NEW YORK BROOKLYN MICROWAVE RESE--ETC F/6 20/3
PROGRESS REPORT NUMBER 41, 15 SEPTEMBER 1975 THROUGH 14 SEPTEMB--ETC(U)
NOV 76 A A OLINER F44620-74-C-0056
UNCLASSIFIED POLY-MRI-452.41-76 NL

3 OF 7
AD
A034764





(a) Sputter-etched before vacuum deposition. Scale: Vertical 0.2 ma/div, Horizontal 1.0 v/div.



(b) Vacuum deposition only. Scale: Vertical 0.2 ma/div, Horizontal 1.0 v/div.

Fig. 2. I-V characteristics of processed wafers.

Using this fabrication technique, unipolar silicon devices were made for further investigation. See report "Avalanche Initiated Oscillation in Unipolar Structures under Pulse Excitation" by B. Senitzky and S. Gottfried in this report.

Joint Services Technical Advisory Committee
F44620-74-C-0056

M. Eschwei and S. Gottfried

REFERENCES

1. J.H. Leck and F. Wilhelmsen, "Effect of Oxygen Absorption on Silicon Surface Conductivity," J. Vac. Sci. Technol., 7, No. 1, 39 (1970).
2. J.M. Andrews, "The Role of Metal-Semiconductor Interface in Silicon Integrated Circuit Technology," J. Vac. Sci. Technol., 11, No. 6, 974 (1974).
3. J.O. McCaldin, "Atom Movement Occurring at Solid Metal-Semiconductor Interfaces," J. Vac. Sci. Technol., 11, No. 6, 994 (1974).
4. T.E. Gallon, M. Prutton and L. Wray, "The (111) Surface of n-Type Silicon," J. Vac. Sci. Technol., 9, No. 2, 911 (1972).
5. R. Glang, "Materials and Processes for Passive Thin Film Components," J. Vac. Sci. Technol., 3, No. 2, 37 (1966).
6. M.P. Lepselter, Bell System Technical Journal, XLV, No. 2, 233 (1966).
7. W.H. Legat, K.H. Tiefert, W.C. Rosveld, "Contact and Interconnection Technology," J. Vac. Sci. Technol., 7, No. 1, 133 (1970).
8. G.J. Kominiak and J.E. Uhl, "In Situ Investigation of Substrate Surface Recontamination During Glow Discharge Sputter Cleaning," J. Vac. Sci. Technol., 13, No. 1, 170 (1976).
9. F.J. Biondi, "Transistor Technology," 3 (D. VanNostrand Company, Princeton, New Jersey, 1958).
10. K.L. Chopra, "Thin Film Phenomena," 319 (McGraw Hill, New York, 1969).
11. S. Tolansky, "Surface Microphotography," 62 (Interscience Publishers, Inc., New York, 1960).
12. Taylor-Hobson, Engis Eastern Corporation, Hasbrouck Heights, New Jersey.
13. Transene Company, Inc., Rowley, Massachusetts.
14. J. Strong, Review Sci. Instru., 6, 97 (1935).
15. B.N. Chapman, "Thin Film Adhesion," J. Vac. Sci. Technol., 11, No. 1, 106 (1974).
16. L.I. Maissel and R. Glang, "Handbook of Thin Film Technology," 12-7 (McGraw Hill, New York, 1970).

NEW SOLID STATE MATERIALS

E. Banks, M. Greenblatt, B.R. McGarvey, S. Nakajima, M. Shone and G. Torre

Research in the area of new materials is devoted to synthesis, crystal growth and characterization of materials which have interesting electrical, optical and magnetic properties, which may provide a basis for new optical, electro-optic and magneto-optic devices. Section A describes the continuation of studies of complex transition metal fluorides. Section B briefly discusses studies of excitation of infrared-visible conversion in $CdF_2(Er^{3+}, Yb^{3+})$. NMR studies of Er-F interactions are described, which will be useful in determining the mechanism of energy transfer in these crystals.

A. Complex Transition Metal Fluorides

The research in this section involves the investigation of ternary fluorides of alkali metals and transition metals, of the general formula $A_xM^{II}_xM^{III}_{1-x}F_3$ (A = K, Rb, Cs; $M^{II} = Fe^{++}$, Mn^{++}, Cr^{II}, etc., $M^{III} = Fe^{+++}$, Cr^{+++}, V^{+++}, etc.). The compounds have the "tetragonal tungsten bronze" structure for x between 0.4 and 0.6, and we have focused on this region because of the well-known ferroelectricity and non-linear optic behavior of oxides having this structure, the possibility of studying ferromagnetic and antiferromagnetic exchange interactions in a complex structure, and the possibility that conductive phases may develop in such systems when mixed valence states of the same element are present. The latter has been ruled out for the cases of Fe, Cr and V, leaving Ti as the only remaining possibility. Previous work has shown antiferromagnetic behavior in most of the phases, except for one case which appears to be ferrimagnetic below 80K. Mössbauer studies of $K_{0.5}FeF_3$ showed evidence of preferential substitution of Fe^{2+} in one of the sites in the unit cell.

A Mössbauer study of selected compositions in tetragonal bronze phases was undertaken to determine the extent of preferential filling of the (2c) sites relative to the (8j) sites. Previous evidence, in $K_{0.5}FeF_3$, showed the Fe^{2+} spectrum to be broadened relative to the Fe^{3+} spectrum, suggesting that the Fe^{2+} ions preferentially filled the (2c) sites while the Fe^{3+} and the remaining Fe^{2+} ions were randomly distributed over the (8j) sites.

A series of samples were prepared of composition $x = 0.5$, some chosen to have only Fe^{3+}, others only Fe^{2+} and others to have just enough Fe^{2+} to fill all the (2c) sites if there were an absolute preference for these sites by Fe^{2+} in competition with all other divalent and trivalent ions in the sample. Mössbauer spectra of all samples were taken at room temperature. One sample, $KFe_{0.4}Mn_{0.6}CrF_6$ was also run at liquid nitrogen temperature. This spectrum showed a peak due to Fe^{3+}, probably due to some FeF_3 in the starting material. Other samples prepared to contain only Fe^{2+} also showed

the presence of Fe^{3+} as additional absorption in the region of the low velocity member of the Fe^{2+} doublet. This necessitated the calculation of area ratios based on the high velocity component only. This procedure is valid in the absence of strong anisotropy of the quadrupole splitting (Goldanskii-Karyagin effect). Areas of peaks and Mössbauer parameters were computer-fitted to the observed spectra. The Fe^{2+} samples showed partial resolution of the peaks due to the (2c) and (8j) sites, which made is possible to calculate area ratios to be compared to predictions based on random occupancy compared to preferential occupancy.

The results of this calculation are shown below. The area ratios are expressed in terms of the ratio of the population of Fe^{2+} in the more abundant (8j) to the less abundant (2c) site.

Composition	Ratio of Fe^{2+} Areas	Numerical Ratio
$KFeVF_6$	0.28/0.19	1.47
$KFeCrF_6$	0.25/0.18	1.40
$KFe_{0.4}Mg_{0.6}VF_6$	0.25/0.20	1.25
$KFe_{0.4}Mn_{0.6}CrF_6$	0.20/0.16	1.25
$KFe_{0.4}Mn_{0.6}Fe^{3+}F_6$	0.12/0.08	1.5

If Fe^{2+} were randomly substituted in the two sites, the ratio expected would be 4:1, the ratio of the abundance of available sites. If the Fe^{2+} had 100% preference for the (2c) sites, the ratio would be 3:2 for samples containing one Fe^{2+} per formula unit and zero for those that contain 0.4 Fe^{2+} per unit, as this would be the amount needed to fill all the (2c) sites. The observed ratios approach the value of 3:2 for samples where one Fe^{2+} is present, and even in samples where 0.4 Fe^{2+} is present. It is significant that the preference does not appear due to the dipositive charge, or ionic radius, as competing divalent ions, Mg^{++} and Mn^{++} were present. Perhaps the (2c) site is less symmetrical, or is more easily distorted by the unsymmetrical electronic distribution of the Fe^{2+} ion, both Mg^{++} and Mn^{++} being spherically symmetrical.

B. Upconversion in CdF_2:YbF_3:ErF_3

A paper summarizing work to date in this system has been prepared for publication in the Journal of the Electrochemical Society. This manuscript summarizes results described in many of our earlier reports on this system.

Recently, Professor Greenblatt has been measuring excitation spectra at Rutgers University. These should permit a study of the relative importance of the "dimers"

$(R.E.^{3+}\text{-}F_i)_2$ discovered in our NMR studies, as compared to isolated Yb^{3+} ions, in contributing to the efficiency of the upconversion process. We are preparing some mixed Yb-Er doped samples for NMR studies to see if mixed pairs ($Yb^{3+}\text{-}Er^{3+}\text{-}F_{i_2}$) are formed and if their number is enhanced by the substitution of Ca^{++} for Cd^{++} in the crystal.

C. Rare Earth-Fluorine Interactions

Previous work in this area has concerned ligand hyperfine interactions between rare earth ions and fluoride ions in CdF_2. One result has been the detection of "dimers" involving rare earth ions, as outlined in the previous section. The following is an abstract of a paper on the ^{19}F NMR study of a single crystal of ErF_3. The crystal was grown by Czochralski pulling of an ErF_3 melt in an HF atmosphere by M. Robinson of the Hughes Research Laboratories and we acknowledge his contribution with gratitude.

The ^{19}F NMR of ErF_3 and $LiErF_4$ powders and a single crystal of ErF_3 has been measured. For ErF_3 the shift tensor was determined for the two different fluoride ion sites in the crystal lattice and the orientation of the principal axes in each site were found. Further, the temperature dependence (200-400°K) of the isotropic and traceless shift tensors were separately determined for each site. The traceless tensor agrees with that calculated from a dipolar model within 10% and its temperature dependence is identical with that reported for the paramagnetic susceptibility of ErF_3. The isotropic shifts for the two sites are found to have different temperature dependences showing that theories of the isotropic shift which assume one hyperfine parameter for all states of the J manifold cannot be applied to this system; analysis of the shift using recent theories show (1) the shift contributions of more than one nearest neighbor rare earth ion are additive and (2) both the covalent and polarization mechanisms of spin transfer must make comparable contributions to account for the results. The analysis of hyperfine parameters reported for the isoelectronic Ho^{2+} in CaF_2 arrives at a similar conclusion about the contributions of covalent and polarization mechanisms for that ion.

It is shown that useful information about the shift tensor can be obtained from powder spectra if there is only one fluoride ion site, as in the case for $LiErF_4$, but not from the rare earth trifluorides which have two different sites in the unit cell.

The full paper has been submitted to the Journal of Magnetic Resonance.

U.S. Army Research Office
DAAG-29-75-G-0096

E. Banks

A SIMPLE NEW TECHNIQUE FOR THE MEASUREMENT OF THE ELASTIC AND MAGNETOELASTIC PROPERTIES OF FERRITES

L.M. Silber

Because of fundamental scientific interest, and in view of the potential applications of ferrites in devices such as variable magnetoelastic delay lines or non-reciprocal elastic surface wave devices, a number of techniques have been developed to measure the fundamental magnetic and elastic properties which determine the behavior of these devices.[1] Among the most important of these properties, when considering the choice of material for device application, are the magnetoelastic coupling coefficient and the elastic losses. While the techniques previously developed can yield precise results for these quantities, these techniques often require elaborate specimen preparation, or construction of special equipment. We propose a simple experimental technique using readily available equipment. It is potentially capable of yielding detailed quantitative information, and should be most convenient as a quick qualitative method of choosing materials having sufficiently interesting properties to warrant investigation by more elaborate techniques.

The experiment is essentially a modification of a technique developed by LeCraw.[2] A small sphere of ferrite is placed in the inductance of a marginal oscillator. (In our experiments a commercial nuclear magnetic resonance gaussmeter was employed, with the hydrogen-lithium sample replaced by the ferrite sphere resting in the bottom of a thin walled glass tube.) A biasing field sufficient to magnetize the sample to saturation as a single domain is applied (i.e., to remove domain walls). The frequency of the marginal oscillator is swept in the usual manner until a resonance is observed. At the frequency corresponding to an acoustic resonance of the sphere, magnetoelastic coupling will excite acoustic oscillations in the sphere. The absorption at resonance is determined by the geometry of the system, the magnitude of the magnetoelastic constants, and the acoustic losses. An equivalent circuit is shown in Figure 1. L represents the

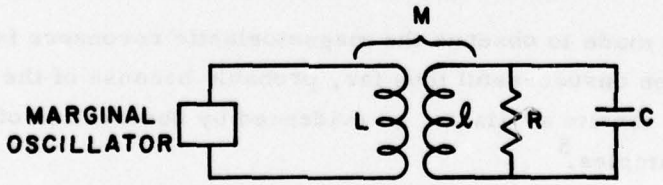


Fig. 1. Equivalent circuit for determining magnetoelastic properties of ferrite sample.

inductance of the oscillator. The mutual inductance M is determined by the geometry of the coil and sample, and the appropriate magnetoelastic coupling coefficients. ℓ, c, and R reflect the elastic properties of the ferrite sample. Assuming the elastic losses are reasonably small, the frequency of resonance is determined by the elastic constants, while the Q of the resonance is determined by the acoustic losses Observation of the strength of the absorption, and absorption line shape allows one to estimate relative values of magnetoelastic coupling coefficients and elastic losses.

To make the results quantitative one would have to calculate the value of M. This value would be different for each of the acoustical modes, because the elastic displacement is different for each mode. For qualitative results, however, one can use a sample with known magnetoelastic coupling coefficient, such as yttrium iron garnet, as a calibrating standard. If the acoustic losses are sufficiently low, they could be measured by pulsing the marginal oscillator and looking at the decay in the acoustic absorption.

Qualitative measurements have been made on single-crystal samples of yttrium iron garnet, gallium-substituted YIG, magnesium ferrite, and a series of nickel-zinc ferrites, all at room temperature. Samples of 1/2 mm to 2 mm in diameter were used. In each instance it was possible to identify several of the acoustic modes of resonance. The results are consistent with those obtained by other techniques. YIG has a small magnetoelastic coupling coefficient, and very small acoustic losses. The magnetoelastic coupling coefficient of nickel-zinc ferrite is much larger, and the acoustic losses much larger. The magnesium ferrite has magnetoelastic coupling coefficient and elastic losses larger than YIG, and smaller than the nickel-zinc ferrites. As reported by LeCraw,[3] one observes two or three closely spaced resonances, rather than a single resonance, corresponding to each mode. They are typically 1/2% to 1% apart. The resonant frequency of each resonance increases slightly with magnetic field, approaching a limit at high fields, due to the magnetoelastic or spin wave contribution to the elastic modulus. It is possible to calculate the appropriate magnetoelastic coupling coefficient from this variation,[4] though this has not yet been done.

An attempt was made to observe the magnetoelastic resonance in samples of cobalt ferrite. This has been unsuccessful thus far, probably because of the poor quality of the samples of cobalt ferrite available, as evidenced by observation of ferromagnetic resonance in these samples.[5]

In summary, the proposed technique offers a convenient method of obtaining qualitative results with readily available commercially-built equipment, and might be useful for a preliminary investigation of a large class of materials to ascertain if more

elaborate experiments are worth undertaking. Measurements as a function of temperature could be readily made. With refinement of the theoretical analysis, it could be developed into a quantitative method of great versatility and convenience.

Joint Services Technical Advisory Committee
F44620-74-C-0056

L.M. Silber

REFERENCES

1. An excellent survey of both experimental techniques and results is presented by R.C. LeCraw and R.L. Comstock in Chapt. 4 of "Physical Acoustics" III, Part B (W.P. Mason, Editor, Academic Press, 1965).
2. R.C. LeCraw and R.L. Comstock, "Physical Acoustics," III, Part B, 169 (W.P. Mason, Editor, Academic Press, 1965).
3. R.C. LeCraw and R.L. Comstock, "Physical Acoustics," III, Part B, 168 (W.P. Mason, Editor, Academic Press, 1965).
4. R.L. Comstock, Proc. IEEE, 53, 1551 (October 1965).
5. L. Assadourian and L.M. Silber, Paper 5E-5 presented at 21st Annual Conference on Magnetism and Magnetic Materials, Philadelphia, Pa., December 1975. Published in Proceedings of the Conference.

MAGNETOCONDUCTIVITY MODULATION OF PERMALLOY FILMS BY SUBSTRATE ELECTROSTRICTION

P. Mazumdar and H. J. Juretschke

The mechanism developed in the preceding report,[1] whereby a thin metallic film forming a plate of a condenser undergoes a change in conductance when the condenser is charged, as a result of electrostrictive deformation of the dielectric, applies equally well to ferromagnetic metals. In addition, however, ferromagnetic metals offer another coupling of strain to film conductivity via the film magnetization M, which leads to some new effects.

In connection with metallic field effect (MFE) measurements on permalloy films we have observed a second harmonic response that is sensitive to both the direction and the magnitude of an applied magnetic field acting in the film plane.[2] Figure 1 shows representative data obtained on a 540 Å 90-100 permalloy film. The rms conductance per (unit surface charge density)^2, $\delta\Sigma(2\omega)$, consists of a magnetic field and angle-independent term (the dashed lines) on which there is superposed a signal varying roughly like 4θ and increasing with decreasing magnetic field like $1/H$.

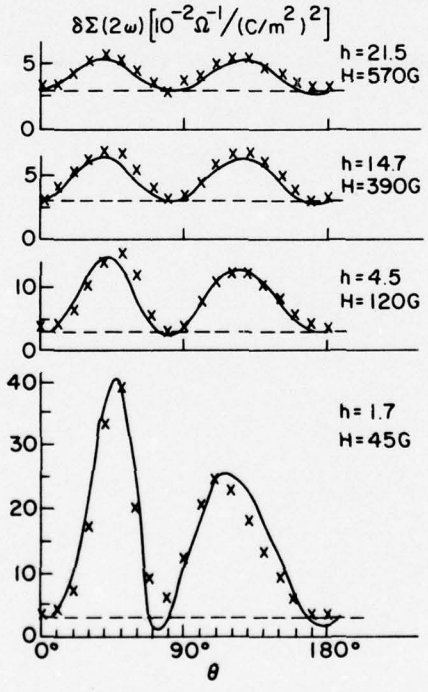


Fig. 1. RMS second harmonic conductance of a permalloy film on mica subjected to a surface charge density q sin ωt. Crosses are experimental points. The solid curves are the fit with the angular function $F(\theta, h)$ of Eq. (4), with $\theta_e = -15^\circ$, at four values of h.

This second harmonic signal is known to be proportional to film thickness, and hence it must be caused by a bulk effect throughout the thickness of the thin film electrode, and initiated by the strain transmitted to the film from the dielectric.

The well-known dependence of the conductivity σ of ferromagnetic metals on the angle θ_m which the magnetization M makes with the current,[3]

$$\sigma = \sigma_0 - \Delta\sigma \cos 2\theta_m \tag{1}$$

contains an explicit 2θ variation. If either σ_0 or $\Delta\sigma$ is affected by isotropic strain, we obtain largely the response discussed in the preceding report, with at most also a 2θ variation. The origin of the 4θ dependence lies in the effect of strain on the angle θ_m of the magnetization in a fixed external magnetic field H.

In a typical magnetically uniaxial permalloy film, the equilibrium direction of M in a prescribed H is given by[4]

$$HM \sin(\theta - \theta_m) = K \sin 2(\theta_m - \theta_e) \tag{2}$$

where θ and θ_m are the direction of H and M in the plane of the film, θ_e is the direction of easy magnetization, and K measures the anisotropy energy. Variations of K with strain will, using Eq. (2), produce variations in θ_m. If we assume a strain variation of K of the form

$$\frac{\delta K}{K} = \sum_i \eta_i e_i \tag{3}$$

then for a polycrystalline permalloy film we derive a change in conductance caused by the modulation of θ_m given by

$$\frac{\delta \Sigma}{\Sigma} = - \frac{\sin 2\theta_m \sin 2(\theta_m - \theta_e)}{2h \cos(\theta - \theta_m) + 2 \cos 2(\theta_m - \theta_e)} \cdot$$

$$\frac{2\Delta\sigma}{\sigma_0} (\eta_1 + \eta_2 - 2 \frac{\sigma_p}{1 - \sigma_p} \eta_3)(s_{11} + s_{12} - s_{13}) \frac{q^2}{2\epsilon_{33}} \tag{4}$$

The second line of Eq. (4) predicts the overall magnitude of the effect in terms of the magnetoconductive response, the effective strain coefficient of anisotropy energy, and the strain produced by a field $E = q/\epsilon_{33}$ in the mica dielectric. The first line describes the angular response $F(\theta, h)$ using the reduced field $h = HM/2K$ and the angle θ. For given θ_e and θ, θ_m is known through Equation (2).

The angular factor $F(\theta, h)$ predicted by Eq. (4) has the prerequisite 4θ variation, and the H dependence of its denominator reproduces well the observed behavior. The

solid curves in Fig. 1 represent the predictions, at four different magnetic fields, of Eq. (4) (slightly modified to include a magnetic contribution to the anisotropy energy) based on a single set of parameters fixed by independent measurements. The absolute magnitude of these curves leads to the experimental description

$$\frac{\delta\Sigma - \delta\Sigma_\infty}{\Sigma} = -6.7\, F(\theta, h)\, q^2 \tag{5}$$

In order to compare this result with the theory of Eq. (4), we must know all the factors of the second line of that equation. The pertinent properties of mica are in the literature[5]

$$\begin{aligned} (s_{11} + s_{12} - s_{13}) &= 5.94 \cdot 10^{-12}\, m^2/\text{newton} \\ \varepsilon_{33} &= 5.5 \cdot 10^{-11}\, \text{farad/m} \end{aligned} \tag{6}$$

The remaining factors were obtained by direct measurements on the sample in question. From the observed magnetoconductivity relation of Eq. (1) we obtain

$$\frac{2\Delta\sigma}{\sigma_0} = 4.6\, 10^{-2} \tag{7}$$

and the strain coefficients of the anisotropy energy were measured under static strains[2] to be

$$(\eta_1 + \eta_2 - 2\,\frac{\sigma_p}{1 - \sigma_p}\,\eta_3) = 3.5 \cdot 10^3 \tag{8}$$

Combining the results of Eqs. (6), (7) and (8), the predicted numerical factor of Eq. (5) is 8.7. Considering the number of interactions involved in the overall description, the agreement between experiment and theory is highly satisfactory. In fact, for other samples the match is even closer. Furthermore, the magnetostrictive properties embodied in Eq. (8) may change sign. We have verified that under those conditions the experimentally observed θ-dependence also reverses. As a final check of the proposed mechanism, the permalloy samples were rigidly glued between thick glass slides, in order to suppress the lateral strains of the mica substrate. The second harmonic signal decreased by more than 80 percent under this treatment.

The overall modulation of the direction of magnetization represented by the data of Fig. 1 is of the order of a second of arc, and is obtained by strains of order 10^{-7}. These small variations offer a new method of analyzing the magnetic equilibrium and the possible spatial distribution of magnetic anisotropy energies in considerable detail. In low fields, the coupling between magnetization and strain demonstrated here can be used to detect small changes in strain, or small changes in the direction of magnetization.

National Science Foundation
DMR 73-07501

P. Mazumdar and H.J. Juretschke

REFERENCES

1. D. Lischner and H.J. Juretschke, "Modulation of Thin Film Conductivity by Substrate Electrostriction," Progress Report No. 41 to JSTAC, Polytech. Inst. of New York, Report No. R-452.41-76 (1976).

2. P. Mazumdar, Ph.D. Thesis, Polytech. Inst. of New York (1976).

3. J.P. Jan, "Galvanomagnetic Effects in Metals," in Solid State Physics, F. Seitz and D. Turnbull, editors, (New York: Academic Press, 1957).

4. M. Prutton, "Thin Ferromagnetic Films," (London: Butterworth and Co., 1964).

5. G. Simmons and H. Wang, "Single Crystal Elastic Constants," (Cambridge, Mass: MIT Press, 1971).

MODULATION OF THIN FILM CONDUCTIVITY BY SUBSTRATE ELECTROSTRICTION

D. Lischner and H.J. Juretschke

In the metallic field effect (MFE) the conductance Σ of an electrode of a parallel plate condenser is altered when the condenser is charged.[1] Experiments on condensers fabricated from thin sheets of muscovite mica with very thin evaporated electrodes give a change of conductance with a.c. charging that has components at both the driving frequency and at the second harmonic.[2] Recent measurements show that most or all of the second harmonic signal is proportional to the electrode thickness,[3] for very thin electrodes, indicating that this part of the response is not a true interface effect. This report presents evidence that in most MFE experiments this second harmonic can be traced to a conductivity modulation of the electrode through strains induced by electrostrictive deformation of the dielectric.

When an electric field E is applied to mica in the direction (001) normal to the cleavage plane, the mica develops isotropic strains in the cleavage plane derivable from the Maxwell stress tensor and the elastic constants of mica as

$$e_1 = e_2 = \frac{1}{2}(s_{11} + s_{12} - s_{13})\varepsilon_{33}E^2 \tag{1}$$

As the change in electrode conductivity is normally linear in the strains, the E^2 dependence of these strains implies a second harmonic response. In order that this effect explain the observed second harmonic it must have the correct magnitude.

The linear strain dependence of the conductivity with the current along the x-direction is a function of all six strains[5]

$$\frac{\delta\sigma}{\sigma} = \sum_j \gamma_{ij} e_j \tag{2}$$

where the γ_{ij} tensor is defined in the coordinate system having x and y in the electrode plane. For electrodes of epitaxial silver or antimony on mica, Eqs. (1) and (2) combine to yield an observable change of conductance $\delta\Sigma$

$$\frac{\delta\Sigma}{\Sigma} = \left[\gamma_{11} + \gamma_{12} - \frac{2c_{13}}{c_{33}}(1 + \gamma_{13})\right] e_1 \tag{3}$$

where the c_{ij} are components of the elastic stiffness tensor of the electrode in the same coordinate system.

The tensor γ_{ij} is only very incompletely known in most metals, if at all. Furthermore, its dependence on electrode thickness or other sample parameters is uncertain. It therefore becomes necessary to determine the combination of the γ_{ij}'s entering into

Eq. (3) by an independent measurement on the same samples used for the MFE.

If a substrate bearing a thin film is bent statically into an arc of a circle, the change of conductance of the film with strain depends on the direction of the plane of the circle relative to the current direction; however, no combination of such measurements will specify γ_{11}, γ_{12} and γ_{13} separately. Nevertheless, it turns out that a major applied strain at 45 degrees to the current, or the sum of the effects at 0° and 90° to the current, precisely produces a $\delta\Sigma$ proportional to the desired combination of γ_{ij} of Equation (3).

The results of some such static measurements is shown in the first column of Table I. These numbers are of the same magnitude as those in the literature for both bulk and thin film samples,[6] but differ in some details. The second column in Table I gives the predicted value of Eq. (3), using the strain of Eq. (1) (normalized to a surface change density of 1 coul/m^2) and the third column lists the experimentally observed second harmonic response of the MFE. The agreement between prediction and experiment is very satisfactory, and indicates that we have identified most, if not all, of the source of the second harmonic contribution of the MFE.

TABLE I. Strain dependence of conductivity, and conductance modulation by electrostriction in silver and antimony.

	$\gamma_{11} + \gamma_{12} - \frac{2c_{13}}{c_{33}}(1+\gamma_{13})$	$\frac{1}{\Sigma}\frac{\delta\Sigma}{\delta q^2}(2\omega)$ $[c/m^2]^{-2}$	
		predicted	measured
Sb	-20.9	-0.77	-0.69
Ag	- 8.66	-0.32	-0.22

As an additional confirmation of the proposed mechanism, the MFE sample was rigidly sandwiched between thick glass slides which effectively suppressed the lateral strains of the mica and its film electrodes. Substantially all of the second harmonic signal could be eliminated by this procedure (although the first harmonic remained completely unaffected).

Beyond successfully identifying a signal which probably accompanies all MFE measurements, the observed coupling between dielectric deformation and electrode conductivity of a parallel plate condenser has other possible uses. It offers a method for detecting strains of the order of 10^{-7} or less, and depending on whether the properties of the electrode or the dielectric are unknown, it can be used to obtain information on γ_{ij} of the electrode material or to determine the strains of the dielectric. This

last case is of particular interest where one looks for the electrostrictive behavior of the dielectric resulting from the strain dependence of its dielectric constants.[7] The agreements of Table I indicate that in mica this contribution can be neglected, although in other dielectrics it can add substantially to the magnitude of Equation (1).

National Science Foundation
DMR 73-07501

D. Lischner and H.J. Juretschke

REFERENCES

1. A. Berman and H.J. Juretschke, "Origin of the Metallic Field Effect," Phys. Rev., B 11, 2903 (1975).
2. H.J. Juretschke and L. Goldstein, "Nonlinear Response in the Metallic Field Effect," Phys. Rev. Letters, 29, 767 (1972).
3. D. Lischner (unpublished).
4. H.J. Juretschke, "Notes on Electrostriction," (unpublished).
5. M.J. Druyvesteyn, "The Variation of the Resistivity of Some Metals with Elastic Deformation," Physica, 17, 748 (1951).
6. R.L. Parker and A. Krinsky, "Electrical Resistance - Strain Characteristics of Thin Evaporated Metal Films," J. Appl. Phys., 34, 2700 (1963).

   B.S. Verma and H.J. Juretschke, "Strain Dependence of the Resistivity of Silver Films," J. Appl. Phys., 41, 4732 (1970).
7. I.S. Zheludev, "Physics of Crystalline Dielectrics," 2, Chapter 9 (New York: Plenum Press, 1971).

ON THE ORIGIN OF THE METALLIC FIELD EFFECT

H.J. Juretschke, D. Lischner and P. Mazumdar

In the two preceding reports,[1,2] we have unambiguously identified the second harmonic signal $\delta\Sigma(2\omega)$ observed in metallic field effect (MFE) measurements as a bulk effect arising from the influence of electric field-induced strains of the substrate on the electrical properties of the metal. This third report summarizes the evidence that the first harmonic signal $\delta\Sigma(\omega)$, a true surface effect, results, at least in part, from the scattering of current carriers by localized interfacial strains linearly proportional to the surface change density q, and it proposes a model to account for this mode of interaction.

The evidence for the role of strain in $\delta\Sigma(\omega)$ comes from several sources:

(a) Angular Dependence of $\delta\Sigma(\omega)$ in Ferromagnetic Metals: When a magnetic field H in the plane of the sample makes an angle θ with respect to the current, as discussed in Ref. 2, both $\delta\Sigma(2\omega)$ and $\delta\Sigma(\omega)$ are functions of θ and H. As shown in Fig. 1, both signals have, in fact, a nearly, identical form of the θ- and H-dependence.[3] Since the second harmonic arises from strain modulation of the anisotropy energy, as demonstrated in Ref. 2, the identical functional form of the signal in $\delta\Sigma(\omega)$ argues strongly that here we are again observing the results of strain modulation of the magnetic anisotropy energy, in this case K_s of the surface region of the ferromagnetic film.

As a matter of fact, most data for $\delta\Sigma(\omega)$ and $\delta\Sigma(2\omega)$ can be matched with a common function $F(\theta, h)$ (see Ref. 2) indicating that magnetic surface and volume properties of permalloy films are not very different. In some cases, phase shifts between the first and second harmonic responses require introducing a small modulation of the surface anisotropy direction, but the overall model still holds.

(b) Influence of Static Strain on Ferromagnetic MFE: When the MFE experiment is carried out with the permalloy film under large static strain, not only does the anisotropy energy change, but the amplitudes of the oscillations of $\delta\Sigma(\omega)$ and $\delta\Sigma(2\omega)$, such as shown in Fig. 1, also change, and by a common factor.[3] This effect on $\delta\Sigma(2\omega)$ must be ascribed to a non-linear dependence of the anisotropy energy on strain. Since the first harmonic $\delta\Sigma(\omega)$ follows the same pattern, it must also be connected to strain, and most likely through the same coupling.

(c) MFE on Ferroelectric Substrates: The MFE experiment of silver on mylar substrates gives results that depend on the polarization state of the mylar.[4] The mylar can be polarized in either direction by cooling it from above 150°C in an applied electric field.[5] In going from the nonpolarized to the polarized state of the substrate, $\delta\Sigma(\omega)$ always changes in the direction of becoming more negative, regardless of the direction

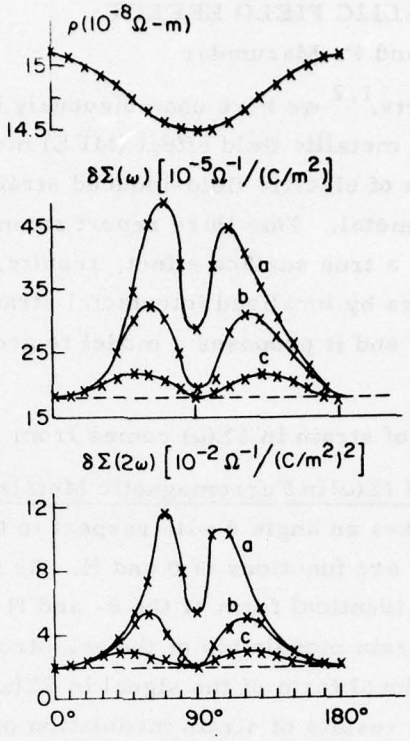


Fig. 1. RMS first and second harmonic conductance changes of a permalloy film on mica subjected to a surface charge density q sin ωt, as a function of the direction of an applied magnetic field in the sample plane. The curves, a,b,c, refer to three fields: H_a = 45 gauss, H_b = 120 gauss, H_c = 530 gauss. For reference, the magneto-resistivity of the sample as a function of angle is also shown.

of polarization. Since the polarized state is accompanied by atomic displacements and strains that are the same for both directions of polarization, the state of strain at the bonded metal-mylar interface is expected to be altered by a large amount under the polarization treatment. Hence changes in both $\delta\Sigma(\omega)$ and strain are related.

(d) The Effect of Large Static Strain on the Non-ferromagnetic MFE: When the substrate is deformed to introduce strains of the order 10^{-4} to 10^{-3} at the interface, $\delta\Sigma(\omega)$ of silver or antimony on mica is observed to change, and, in fact, it often moves irreversibly to a new value.[4] It is likely that under such large strains the interface relieves the local stresses by faults and dislocations, and consequently the state of strain at the undeformed interface is permanently altered under such treatment.

(e) MFE as a Function of Temperature: The first harmonic of silver, gold, or antimony on mica, $\delta\Sigma(\omega)$, changes reversibly with temperature by amounts comparable to or larger than those introduced by large static strains.[6] Because of the method of

sample preparation at high temperature, differential thermal expansion of sample and substrate always produces strains, and these are expected to be most severe in the region of the interface.

In summary, we have observed changes in $\delta\Sigma(\omega)$ only when the external parameter influencing the system can be tied to changes in interfacial strain. On the other hand, as described in Refs. 1 and 2, a macroscopic clamping of the system, while successfully suppressing $\delta\Sigma(2\omega)$ has very little or no effect on $\delta\Sigma(\omega)$. We therefore conclude that any interfacial strain involved in the scattering giving rise to $\delta\Sigma(\omega)$ must be either normal to the interface, since such local strains cannot be clamped externally, or it could also be a microscopic strain distribution in the plane of the interface having a vanishing macroscopic average.

A model having these features can be developed along the lines recently proposed to understand the interaction of freely falling electrons with conducting walls in a gravitational field.[7] If a charge Q occupies a fixed site at the metal-insulator interface, its electric field will exert forces on the unscreened ions in its immediate neighborhood that have two properties:

(1) The forces will change direction with the sign of Q

(2) The net force in the plane of the interface is zero, but the net force in the metal normal to the interface is non-vanishing.

As a result of these localized forces the material in the immediate neighborhood of Q will deform elastically. If each of the charges deposited at the interface behaves independently, we have a model in which these charges alter the local state of strain with a distribution proportional to the surface charge density q. These strain fields will scatter current carriers, apart from the direct scattering by each screened Q.

In the bulk, the predominant effect of local strain on the scattering of current carriers arises from a screened deformation potential linear in the strains.[8] If the surface region is considered to be naturally strained, any additional surface charge density-induced strain will cause a change in surface scattering linear in this strain, and hence is proportional to both q, and to the original strain.

Thus we obtain a model for the MFE which contains the major ingredients of the observed effects. It is similar to one originally proposed in terms of direct scattering by charge patches,[9] except that now the scattering is mediated through local interfacial strains, as required by the observed strain dependence of the various phenomena. A prediction of the order of magnitude of the scattering expected in this model is difficult, since so little is known about either the actual state of strain in the surface or the scattering cross section of the strained region. However, from the known magnitude of the

pressure dependence of the work function of metals,[10] combined with a thermodynamic identity, the interfacial strains e are related to the surface charge density by $e = 10^{-1} q$ (with q in Coul/m^2). If we then assume that scattering with surface strains is similar to that observed in bulk, we obtain the estimate for silver $\delta\Sigma(\omega) = -10^{-1}/(\text{ohm-C/m}^2)$, which is compatible with observation. Alternatively, the model of Ref. 9, taking into account that its quadratic term is masked by the electrostrictive signal,[1] requires a coupling between surface conductance Σ_s and strain e of the type $\delta\Sigma_s = -10\, e^2 (\text{ohm})^{-1}$. A relation of such magnitude is not inconsistent with the scattering strength of locally strained regions.[8] Interestingly, this model also requires a static strain in the surface of order 10^{-2} or larger, which suggests that such strain can account for a significant portion of the overall surface scattering of most metal surfaces.

Joint Services Technical Advisory Committee
F44620-74-C-0056 H.J. Juretschke, D. Lischner and P. Mazumdar

National Science Foundation
DMR 73-07501

REFERENCES

1. D. Lischner and H.J. Juretschke, "Modulation of Thin Film Conductivity by Substrate Electrostriction," Progress Report No. 41 to JSTAC, Polytech. Inst. of New York, Report No. R-452.41-76 (1976).
2. P. Mazumdar and H.J. Juretschke, "Magnetoconductivity Modulation of Permalloy Films by Substrate Electrostriction," Progress Report No. 41 to JSTAC, Polytech. Inst. of New York, Report No. R-452.41-76 (1976).
3. P. Mazumdar, Ph.D. Thesis, Polytech. Inst. of New York (1976).
4. D. Lischner, Ph.D. Thesis, Polytech. Inst. of New York (1976).
5. R.N. Varney and H.T. Hahn, "Electrets and Electrostatic Measurement," Amer. J. Phys., 43, 509 (1975).
6. A. Berman and H.J. Juretschke, "Origin of the Metallic Field Effect," Phys. Rev., B 11, 2903 (1975).
7. C. Herring, "Gravitationally Induced Electric Field Near a Conductor, and its Relation to the Surface-Stress Concept," Phys. Rev. 171, 1361 (1968).
8. J.M. Ziman, "Electrons and Phonons," Oxford U. Press, 228 (1960).
9. H.J. Juretschke and L. Goldstein, "Non-linear Response in the Metallic Field Effect," Phys. Rev. Letters, 29, 767 (1972).
10. K.H. Leners and R.J. Kearney, "Stress-Dependent Contact Potential in Copper," Phys. Rev., B 6, 2943 (1972).

THE INTERACTION OF INTENSE LASER RADIATION WITH METAL SURFACES

M. C. Newstein and N. Solimene

The interaction of intense optical radiation with condensed matter is an active field of study because of the variety of technical applications of the resultant phenomena as well as because of the scientific interest in exploring previously unavailable ranges of the interaction parameters. At intensities of 10^4-10^7 watts/cm^2 the applications include heating, deformation, drilling, cutting, welding, etc. of materials.[1,2,3] At about 10^8 watts/cm^2 plasma formation occurs and there is applicational and diagnostic interest in laser produced plasmas as sources of electromagnetic radiation from the infrared to the X-ray region, as well as of electrons and highly ionized atoms.[4,5,6] Applications to laser fusion become possible at power densities of 10^{15} watts/cm^2 and higher.[7,8]

We have initiated a theoretical study of the interaction with metal surfaces of short optical pulses with intensities in the range from 10^6 to 10^9 watts/cm^2. We are modelling experimental conditions which are such as to allow for the study of physical phenomena not fully treated in previous investigations. In contrast to most of the previous material processing studies, the pulse durations are short enough that equalization of the electron and lattice temperatures may not occur. Furthermore, there may be significant variations of optical properties over the effective skin depth, hence the Fresnel-formula does not properly describe the absorption and reflection properties of the surface, rather this is obtained from a self consistent solution of the field-matter equations for the system. On the other hand, in contrast to most of the fusion studies, our pulses are weak enough that within times of experimental interest the material remains in a condensed phase and the thermal parameters are those descriptive of the degenerate electron gas in metals.

We have constructed a formulation of the problem which treats separately many of the complex-simultaneously occurring physical phenomena in order better to evaluate their individual physical significance. The first class of problems studied eliminated the effect of thermal diffusion and concentrated on the development of numerical methods for obtaining the reflection and absorption coefficients. We assume variation in a one-spatial dimension (plane wave input). The metal is assumed to occupy the space described by $z > 0$. It is characterized by the conduction electron plasma frequency, ω_p, electron collision frequency for momentum exchange, ν_p, and for energy exchange, ν_E, as well as the electron and lattice temperatures, T_e and T_L. The collision frequencies are sums of temperature independent terms (due to the effects of extrinsic lattice imperfections) as well as temperature dependent terms (due to electron-phonon collisions). In general, these parameters are functions of space and time, some of this dependence

is via the local electric field strength. We assume that the field (of frequency ω) within the metal ($z > 0$) is of the form

$$E = \mathrm{Re}\ \mathcal{E}(z,t)\, e^{-i\omega t} \tag{1}$$

$$H = \mathrm{Re}\ \mathcal{H}(z,t)\, e^{-i\omega t} \tag{2}$$

where the time dependence of the complex envelopes $\mathcal{E}$ and $\mathcal{H}$ is slow compared to an optical period $2\pi/\omega$. With this approximation, and further assumptions, to be discussed, the dimensionless form of the field-matter equations becomes:

$$\frac{\partial \mathcal{E}}{\partial z} = -i\omega \mathcal{H} \tag{3}$$

$$\frac{\partial \mathcal{H}}{\partial z} = -i\omega n^2 \mathcal{E} \tag{4}$$

$$n^2 = 1 + \frac{i/\omega}{\nu - i\omega} \tag{5}$$

$$\nu_p = KT_L(z,t) \tag{6}$$

$$\frac{\partial U_e}{\partial t} = \left(\frac{1}{4.8 \times 10^7}\right) \frac{\nu_p |\mathcal{E}|^2}{\nu_p^2 + \omega^2} - \nu_E \frac{T_e}{T_F}\left[T_e - T_L\right] \tag{7}$$

$$\frac{\partial T_L}{\partial t} = +\nu_E \frac{T_e}{T_F}\left[T_e - T_L\right] \tag{8}$$

$$U_e(T_e) = \frac{[T_e^2]}{T_e + 2T_F} \tag{9}$$

In the above, $\mathcal{E}$ is measured in units of $(|e|/\omega_p mc)^{-1}$; $\mathcal{H}$ in $(\varepsilon_0 |e|/\omega_p mc^2)^{-1}$, ν in ω_p, t in $1/\omega_p$ and z in c/ω_p. The electron, lattice and Fermi temperatures T_e, T_L, and T_F are measured in units of the melting temperature T_M. Equations (3)-(5) are the field equations with the effect of the medium described in terms of the complex index $\sqrt{n^2}$. Equation (6) describes the assumption that the momentum-exchange collision frequency is proportional to the lattice temperature. Equation (7) gives the rate of change of energy per conduction electron due to the competition between heating by the applied field (first term) and relaxation to the lattice temperature (second term). A phenomological treatment of the relaxation process in the kinetic equations leads to this form. Equation (8) describes the lattice heating via electron-lattice energy exchange collisions and Eq. (9) is a simple expression for the electron energy as a function of electron temperature which represents the behavior for temperatures less than and comparable to the Fermi temperature in a form suitable for numerical calculations.

The reflection coefficient is determined by the surface value (z = 0) of the admittance Y, given by

$$Y(z,t) = -\frac{H}{\mathcal{E}} = -\frac{i}{\omega}\frac{\partial}{\partial z}\ln \mathcal{E} \tag{10}$$

This equation implies:

$$\mathcal{E}(z,t) = \mathcal{E}(0,t)\, e^{i\omega \int_0^z dz' Y(z',t)} \tag{11}$$

and Y satisfies

$$\frac{\partial Y}{\partial z} = -i\omega Y^2 + i\omega \overline{n^2}(z,t) \tag{12}$$

with the boundary conditions

$$Y(z=\infty,t) = n(\infty,t) = \sqrt{n^2(\infty, t=0)} \tag{13}$$

When Eqs. (12) and (13) are solved for Y(z,t), the complex field reflection coefficient at each time t is obtained from:

$$R(t) = \frac{1 - Y(0,t)}{1 + Y(0,t)} \tag{14}$$

We report below some results of this analysis for the action of 5000 A laser radiation on an aluminum target with incident intensities of 10^6, 10^7, and 10^8 watts/cm^2. It should be emphasized that we have excluded the effects of thermal diffusion for the reason previously given. For a melting temperature of 933°K, the ratio of the Fermi to the melting temperature is $T_F = 857$. The ratio of the electron-energy relaxation time to momentum relaxation time is not known. For the purposes of the present calculation we have assumed $\nu_E/\nu_p = 10^{-5}$, this is probably an underestimate.

In Fig. 1 we have plotted the surface values of the electron and lattice temperature (relative to the melting temperature) as well as the power reflection coefficient $|R|^2$ as functions of time. Common features of all three cases are the rapid rise and then decay of the electron temperature toward equilibrium with the more slowly rising lattice temperature, and the relatively small variations in the power reflection coefficients.

A main effect of thermal diffusion will be to decrease the local rate of increase of the electron temperature. This part of the study is being implemented.

Joint Services Technical Advisory Committee
F44620-74-C-0056

M.C. Newstein

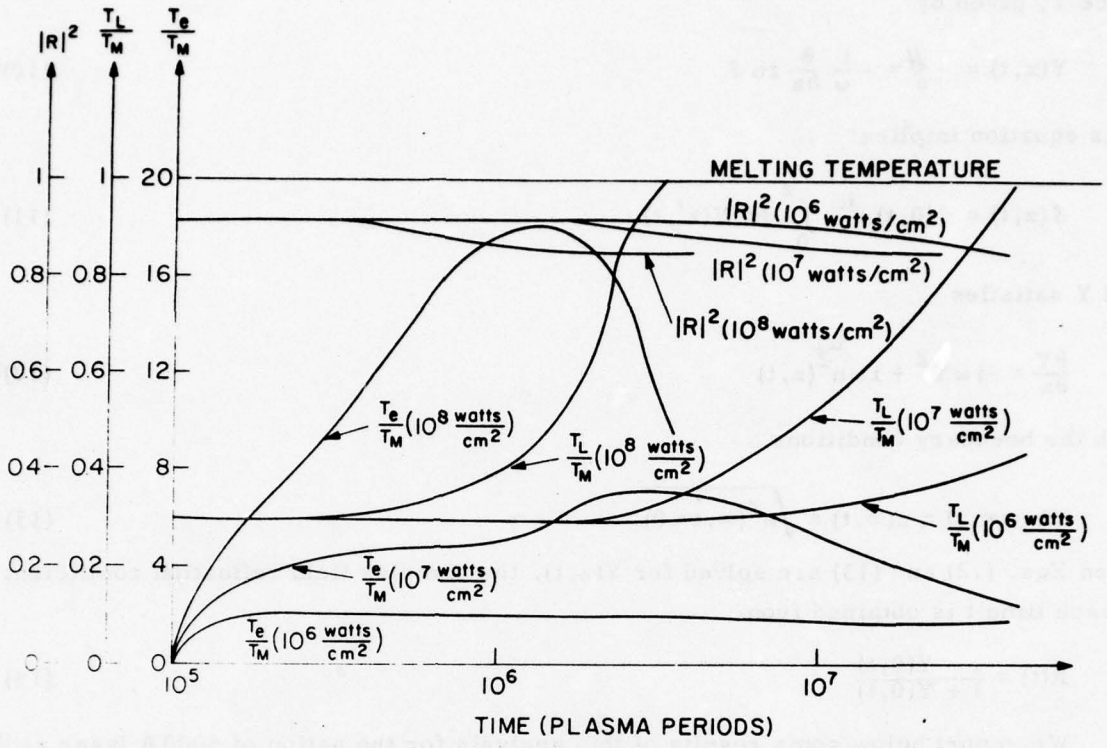


Fig. 1. Illustrating the temporal variation of electron temperature, T_e, lattice temperature, T_L, and intensity reflection coefficient, $|R|^2$ as functions of time at the surface of an aluminum target irradiated with varying intensities of 5000 A laser radiation. The effects of thermal diffusion have been excluded from this study in order to separately evaluate their importance.

REFERENCES

1. I.P. Shkarofsky, "Industrial Applications of High Power Laser Beams," RCA Review, 36, 336 (1975).
2. L.A. Weaver, "Machining and Welding Applications," in Laser Applications, 1, Ed. Monte Ross (New York: Academic Press, 1971).
3. M.I. Cohen, "Material Processing," Sec. F4 in Laser Handbook, Eds. F.T. Arecchi and E.O. Schulz-Dubois (Amsterdam: North-Holland Publishing Co., 1972).
4. For recent reviews of this area see, P. Mulser, R. Sigel and S. Witkowski, "Plasma Production by Laser," Physics Reports, 6, 187 (1973); and
5. O.N. Krokin, "Generation of High Temperature Vapors and Plasmas by Laser Radiation," Sec. E7, Laser Handbook, ibid.
6. D.V. Giovanielli and R.P. Godwin, "Optics in Laser Produced Plasmas," Am. J. Phys., 43, 808 (1975).
7. K.A. Brveckner and S. Jorna, "Laser Driven Fusion," Rev. Mod. Phys., 46, 325 (1974); and
8. F.L. Ribe, "Fusion Reactor Systems," Rev. Mod. Phys., 47, 7 (1975).

SELF-FOCUSING OF COHERENT PULSES

M.C. Newstein and F. Mattar

The physical mechanisms responsible for coherent self focusing[1,2,3] can be elucidated by following the propagation of the transverse energy flow correlated with the corresponding values of the field amplitude. The electric field E is given by

$$E = \mathrm{Re}\ \mathcal{E}\ \exp(i[\omega t - kz]) \tag{1}$$

where the complex amplitude, $\mathcal{E}$ is represented in terms of the magnitude A and phase ϕ by

$$\mathcal{E} = A \exp(i\phi) \tag{2}$$

The longitudinal component of the energy current vector $\overline{J}$ is

$$J_z = A^2 \tag{3}$$

and the transverse part is given by[2]

$$J_T = A^2 \frac{\partial \phi}{\partial \rho} \tag{4}$$

In Fig. 1 the field amplitude is plotted versus the retarded time for three stages of the propagation process: (a) the reshaping region; (b) the build-up region; and (c) the focal region. The transverse energy current is plotted versus the retarded time for the same three distances in Figures 1(d), (e), and (f). In each case the plots are given for several values of the transverse coordinate, ρ. Positive values of the transverse energy flow correspond to outward flow, and negative values to inward. Figure 1 clearly illustrates the following features of the self-focusing process. In the earliest stages of the propagation (Figs. (a) and (d)) the near axis energy current is outward for most of the pulse time, but becomes inward (self-focusing) toward the rear ($\tau \approx 2.4$). For this value of τ the field amplitude (Fig. 1(a)) is already past its peak and has a small value. As we proceed into the reshaping region (Figs. (b) and (e)) the near axis peak amplitude moves back in time (corresponding to the fact that the group velocity is less than c/n) while the temporal location of the change from focusing to defocusing energy flow remains the same. This leads to a large increase in the value of the transverse energy flow. In Fig. 1(c) and (f) (in the focal plane) the peak amplitude occurs at $\tau = 2.4$. The energy current flow in the earlier stages of the pulse is now outgoing corresponding to power which has already focused and is now diverging.

The results of the earlier stage (a) are in quantitative agreement with the analytic predictions of the perturbation theory presented in the last report.[1]

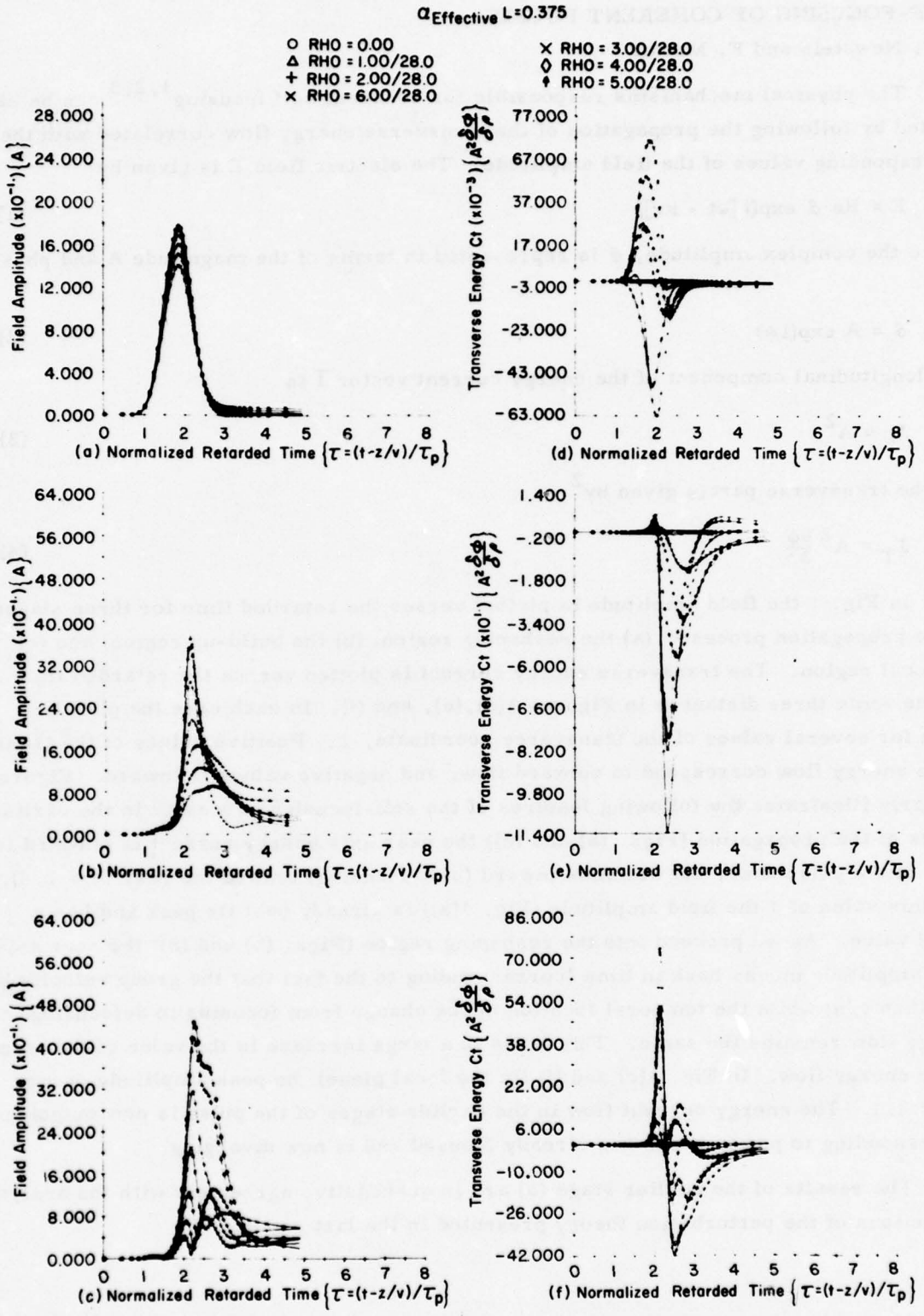


Fig. 1. The field amplitude (a), (b), (c) and the transverse energy current (d), (e), (f) for several radii versus the regarded time for three stages of the propagation: the reshaping region, the build-up region and the focal region (as a function of the transverse coordinate).

Experimental verification of the phenomenon of coherent self focusing has recently been reported.[4,5,6]

Joint Services Technical Advisory Committee
F44620-74-C-0056

M.C. Newstein and F. Mattar

Mobil Oil Corporation

REFERENCES

1. M. Newstein, F. Mattar and D. Ramakrishnan, "Self Focusing of Coherent Pulses," Progress Report No. 40 to JSTAC, Polytech. Inst. of New York, Report No. R-452.40-75, 263-275 (1975).

2. F.P. Mattar and M.C. Newstein, "Self Focusing of Coherent Optical Pulses in Resonant Absorbing Media," Opt. Comm., 18, 70-72 (1976).

3. F.P. Mattar, "Transverse Effects Associated with the Propagation of Coherent Optical Pulses in Resonant Absorbing Media," Ph.D. Dissertation, Polytech. Inst. of New York (June 1976).

4. H.M. Gibbs, B. Bolger and L. Baede, "On Resonance Self-Focusing of Optical Pulses Propagating Coherently in Sodium," Opt. Comm., 18, 199-200 (1976).

5. G. Forster and P.E. Toschek, to be published.

6. H.M. Gibbs, B. Bolger, F.P. Mattar, M.C. Newstein, G. Forster and P.E. Toschek, "Coherent on-Resonance Self-Focusing of Optical Pulses in Absorbers," submitted for publication.

ANALYSIS OF STIMULATED BRILLOUIN SCATTERING IN PLASMAS

E.S. Cassedy

It is well known that intense laser light can cause stimulated Brillouin scattering (SBS) in a variety of transparent media, including sub-critical plasmas. In the case of plasmas, the coupling results from the quiver motion of electrons, as driven by the laser electric field. A phase-synchronous coupling at three frequencies then evolves through the mechanism of V x B forces in the plasma, resulting in the generation of density variations of electrons and ions (the ion-acoustic wave) and the generation of a light wave which is frequency down-shifted from the pumping laser frequency. The phase-synchronous coupling is of the same parametric type as that used to explain the parametric-decay interaction (PDI), only the down-shifted light wave in the SBS process replaces the down-shifted Langmuir wave of the PDI.

Recently it has been suggested[1] that a purely-growing instability exists in conjunction with SBS, developing colinearly along the axis of the laser beam, when the plasma is illuminated at near-critical densities. If this purely-growing instability could be viewed as the analog of the oscillating two-stream instability[2] (OTSI), then we could conclude that it exists as part of a four-frequency interaction involving the up-shifted (or anti-Stokes) light wave, in addition to the down-shifted (Stokes) light, the ion-acoustic wave and the laser-pump light wave. Drake, et al.[3] have considered a formulation including up-shifted and down-shifted light waves, but their results indicate a purely-growing instability transverse to the (laser) pump beam leading to a modulational instability and self-filamentation of the laser beam. The present formulation, by contrast, is colinear where the instability would develop along the pump beam axis and therefore constitute a four-frequency interaction analogous to the OTSI.

Such a formulation for the four-frequency parametric coupling of light waves and the ion-acoustic wave has been derived elsewhere[4] and need not be repeated here. In the Fourier-LaPlace-transformed space, that formulation could be written in the identical form as done previously[5] for the PDI-OTSI case and the dispersion relation expressed in the form:

$$\begin{vmatrix} D_{-1} & 1 & 0 \\ 1 & D_0 & 1 \\ 0 & 1 & D_{+1} \end{vmatrix} = 0 \qquad (1)$$

where in this case, for a traveling-wave pump field $E(x,t) = E_0 \cos(\omega_0 t - k_0 x)$ we have:

$$D_0 = \frac{1}{K}\left[\omega^2 - S_i^2 k^2 - i\,\nu_i \omega\right]$$

$$D_{\pm 1} = \frac{1}{K}\left[(\omega \pm \omega_0)^2 - c^2(k \pm k_0)^2 - \omega_{pe}^2 - i\nu_e(\omega \pm \omega_0)\right]$$

with

$$K = \omega_{pi} k V_0/\sqrt{2} \quad , \quad V_0 = e E_0/m_e\omega_0$$

ω_{pe}, ω_{pi} = plasma electron and ion resonance frequencies, respectively

ν_e, ν_i = electron and ion damping factors

c = velocity of light.

A. The Dispersion Relation

The dispersion relation of Eq. (1) may be used to investigate the four-frequency parametric interaction of light waves and the ion-acoustic wave. Proceeding again as in the previous case,[5] we search for roots to Eq. (1) utilizing Newton's Method on the digital computer and display these solutions as conformal mappings of $\omega(k)$ on the complex-k plane. The mapping for an SBS interaction is shown in Fig. 1 for parameters corresponding to a laser-plasma pumped at 99.5% of the critical frequency. The parameters are all realistic with respect to an experimental plasma, with the exception of the damping terms (ν_i and ν_e) which have been taken as zero. The pumping field strength (E_0) corresponds to an incident laser intensity of about 3×10^{17} W/cm^2.

The Riemann sheet shown is identified with the ion-acoustic mode for this SBS interaction, since the ion-acoustic wave is the basic mode of the interaction. Absolute instability is indicated by the existence of a saddle-point in the mapping. This is to be expected since the damping has been taken as zero and we must therefore be above the absolute-instability threshold[6] for this backward-wave parametric interaction.

These computed results for the SBS may be compared to the results of the coupled-mode approximation, as applied to this interaction. A quadratic dispersion relation for the coupled-mode approximation has previously been derived for generic parametric interactions with finite pump wavelength (i.e., $k_0 > 0$) (Ref. 6) and a resonant cut-off propagation characteristic for one of the coupled modes. The application of these generic coupled-mode results to the case of SBS in a sub-critical plasma yields the following quadratic approximation dispersion relation:

$$(\bar{\omega} - S_i \bar{k})(\bar{\omega} + v_g \bar{k}) = -(S_i + v_g)^2 \frac{A^2}{4} \qquad (2)$$

where the pertubation variables[5,6] are:

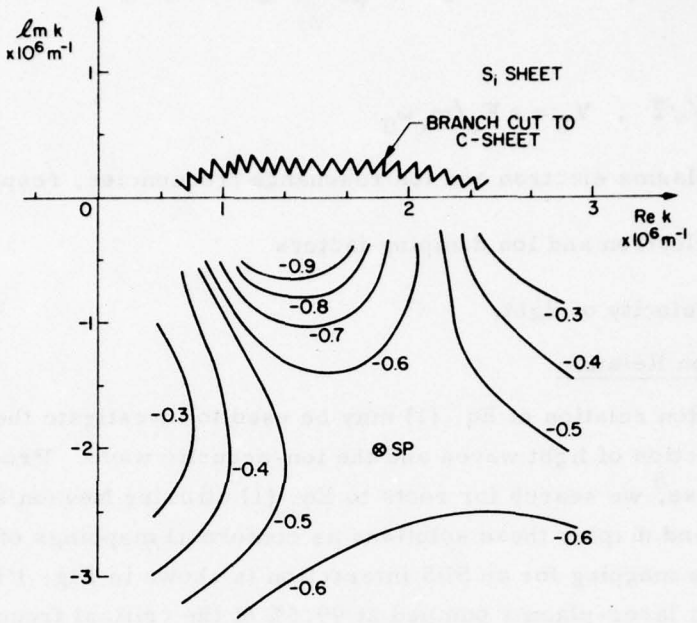


Fig. 1. Stimulated Brillouin scattering - dispersion function.

Parameters

$\nu_i = \nu_e = 0$, $S_i = 2 \times 10^5$ m/sec., $S_e = 1.5 \times 10^7$ m/sec.

$m_i = 3.4 \times 10^{-27}$ Kg, $m_e = 9.1 \times 10^{-31}$ Kg

$\omega_0 = 1.80 \times 10^{15}$ sec.$^{-1}$, $\omega_{pe} = 1.79 \times 10^{15}$ sec.$^{-1}$

$E_0 = 1.5 \times 10^{12}$ volt/m

Derived Factors (quadratic approximation)

$k_I^{SBS} = 1.25 \times 10^6$ m^{-1}, $\omega_I^{SBS} = 0.25 \times 10^{12}$ sec.$^{-1}$

$k_{sp} = (1.25 - i\ 3.53) \times 10^6$ m^{-1}

$\omega(k_{sp}) = (0.25 - i\ 1.42) \times 10^{12}$ sec.$^{-1}$

Estimated Saddle Point

$k_{sp} = (2.2 - i\ 2.0) \times 10^6$ m^{-1}

$\omega(k_{sp}) = (0.60 - i\ 0.57) \times 10^{12}$ sec.$^{-1}$

Legend for Contours: [——— Im $\omega \times 10^{12}$ sec.$^{-1}$]

$$\bar{\omega} = \omega - \omega_I^{SBS}$$

$$\bar{k} = k - k_I^{SBS} \quad ,$$

The interaction wavenumber and frequency[6] are:

$$k_I^{SBS} = \frac{c^2 k_0 - S_i \omega_0 + \sqrt{S_i^2 \omega_{pe}^2 + c^2 (S_i k_i - \omega_0)^2 - c^2 \omega_{pe}^2}}{c^2 - S_i^2} + i \frac{\nu_e - \nu_i}{2(S_i + v_g)}$$

$$\omega_I^{SBS} = S_i k_I^{SBS} + \frac{i \nu_i}{2}$$

the branch-point separation[6,7] is:

$$A = \left[\frac{K^2}{\omega_I^{SBS} (\omega_0 - \omega_I^{SBS}) (S_i + v_g)^2} \right]^{-1/2}$$

and the local group velocity[6,7] is:

$$v_g = \frac{c^2 (k_0 - k_I^{SBS})}{(\omega_0 - \omega_I^{SBS})}$$

This approximate dispersion relation yields a saddle-point of the following parameters:

$$k_{sp} = (1.25 - i\, 3.53) \times 10^6 \, m^{-1}$$

$$\omega(k_{sp}) = (0.25 - i\, 1.42) \times 10^{12} \, sec.^{-1}$$

for the case calculated above, which may be compared with the saddle point estimated from the contour plot of Fig. 1:

$$k_{sp} \simeq (2.2 - i\, 2.0) \times 10^6 \, m^{-1}$$

$$\omega(k_{sp}) \simeq (0.60 - i\, 0.57) \times 10^{12} \, sec.^{-1}$$

Thus, we can observe that the inclusion of the anti-Stokes light component in the dispersion relation results in a quantitative effect on the frequency, growth rate and wavelength of the ion-wave in this SBS interaction in a near-critical ($\omega_{pe} = 0.995 \omega_0$) plasma, but that the essential qualitative features for absolute instability are maintained.

The branch cut shown in the vicinity of the Re-k axis on Fig. 1 indicates the connection from the ion-acoustic sheet to one of the neighboring sheets corresponding to

the Stokes or anti-Stokes shifted light waves. That is, on those sheets the dispersion function would be closely approximated by either of the following:

$$\omega(k) = \begin{cases} \omega_0 - \sqrt{\omega_{ep}^2 + c^2(k - k_0)^2} \\ -\omega_0 + \sqrt{\omega_{ep}^2 + c^2(k + k_0)^2} \end{cases} \tag{3}$$

in any region where parametric interaction was not manifested. These dispersion functions are of a similar form as those discussed for the PDI-OTSI, with the velocity of light (c) here replacing the electron thermal velocity (S_e). In this case, unlike the previous, however, there are wavenumber shifts ($k \mp k_0$) involved concurrently with the frequency shifts ($\omega \mp \omega_0$) for the Stokes and anti-Stokes light components, respectively. These sheets will henceforth be referred to as "c-sheets."

An investigation of the complete dispersion relation of Eq. (1) on the c-sheets displays no solutions which suggest instabilities. That is, no saddle-point regions of the type analogous to that found for the OTSI were found for the $\omega(k)$ conformal mappings on the c-sheets as solutions of Equation (1). Instead, all digital computer solutions showed solutions closely approximating those values calculated by the uncoupled c-sheet solutions (Equation (3)). This suggests, of course, that no interaction between up and down-shifted light waves exists for this formulation and would lead us to conclude that no purely-growing instability for this case, in analogy to the OTSI, exists.

Confirmation of the absence of a purely-growing instability in the presence of SBS has been carried out[8] by analysis of a two-mode approximation similar to that carried out previously for other parametric interactions.[6,7] It should be noted, however, that this absence of an interaction holds only for a traveling-wave pump field (see Equation (1)). If, on the other hand, a standing-wave pump field is assumed, then an absolute instability of the purely-growing type is indicated.[8,9]

National Science Foundation
ENG 73-03774

E.S. Cassedy

REFERENCES

1. J.R. Faehl and W.L. Kruer, Bull. Amer. Phys. Soc., 19, 919 (1974).
2. K. Nishikawa, Phys. Soc. Japan, 24, 1152 (1968); P.K. Kaw and J.M. Dawson, Phys. Fl., 12, 2586 (1966); J.R. Sanmartin, Phys. Fl., 13, 1533 (1970).
3. J.F. Drake, P.K. Kaw, Y.C. Lee, G. Schmidt, C.S. Liu and M.N. Rosenbluth, Phys. Fl., 17, 778 (1974).
4. L.M. Garbunou, Sov. Phys. JETP, 28, 1220 (1969); K.A. Brueckner and S. Jorna, Rev. Mod. Phys., 46, 325 (1974).
5. E. Cassedy, H. Huang and S. Peng, "On the Co-existence of Parametric Decay and Non-oscillatory Instabilities in Plasmas," Progress Report No. 39 to JSTAC, Polytech. Inst. of New York, Report No. R-452.39-74, 240-245 (1974).

6. E.S. Cassedy and C.R. Evans, Ph.D. Dissertation, Polytech. Inst. of New York (1971).

7. E.S. Cassedy and P. Mulser, Phys. Rev. (A), 10, 2349-2359 (1974).

8. E.S. Cassedy, "Purely-growing Parametric Instabilities in Plasmas at Near-Critical Densities," J. Plasma Phys. (British), accepted for publication (1976).

9. J.R. Faehl, E.S. Cassedy and W.L. Kruer, to be presented at the 1976 Meeting Plasma Division, Amer. Phys. Soc., San Francisco, California.

OBSERVATION OF WAVE PACKET BIFURCATION IN A MAGNETO-PLASMA COLUMN

E. E. Kunhardt and B. R. Cheo

A. Background and Experimental Setup

The bifurcation of wavepackets has been predicted by modulation theory of non-linear dispersive waves.[1,2] We report here the first observation of this effect using a magneto-plasma column surrounded by a conducting waveguide. Using Maxwell's equations and a cold plasma model, it can be shown that for weak axial magnetic field the normalized wavepacket amplitude $a(z,t)$ of the lowest propagating plasma mode (TM_{01}) excited by an axial impulse electric field at a point along the column, is governed by the Korteweg-deVries equation:[3]

$$\partial_z a + v_0^{-1} \partial_t a - \beta \partial_t^3 a + (\omega_c b)^{-1} a \, \partial_t a = 0 \tag{1}$$

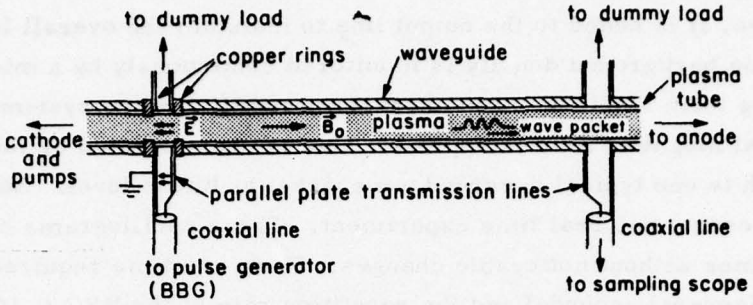
where the longitudinal component of electric field is proportional to $\partial_t a$, ω_c = cyclotron frequency, b = column radius, and $\beta = 1/2 v_0 \omega_c^2$. v_0, the phase velocity of small oscillations in the long wavelength limit, is given by:[4] $v_0 = b\omega_p\omega_c/2.405(\omega_p^2 + \omega_c^2)^{1/2}$. The non-linear dispersion relation is given by:[3]

$$k(\omega) \approx k_0(\omega) + a^2 k_2(\omega) \tag{2}$$

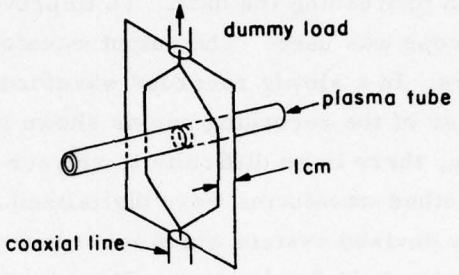
with $k_0(\omega) = \omega/v_0 + \beta\omega^3$ and $k_2(\omega) = -1/24\,\beta\omega_p^2 b^2 \omega$. Since $k_2 k_0'' < 0$, the modulation equations are hyperbolic, and subject to the bifurcation predicted; i.e., there will be two real characteristics and two group velocities:

$$v_g \approx k_0'^{-1} \pm a/2b\omega_p k_0'^2 \tag{3}$$

The experimental setup is schematically shown in Fig. 1(a). The wavepackets are excited by a homemade device called a bouncing ball generator (BBG) as used by Pleshko and Palocz[5] in their experiments to observe the Brillouin and Sommerfeld precursors. Here a modified version is used which is capable of producing pulses with peak voltages of 3.2 KV (at 50 Ω) with a duration approximately 400 ps. This high intensity ultra-short pulse is applied axially along the plasma column through the launcher at the input point. The initial disturbance caused by the impulsive field evolves into a wavepacket as it propagates down the column. The plasma is the positive column of an argon discharge 85 cm long with an axial B field up to 1.2 KG. The plasma background neutral density $\sim 3.3 \times 10^{13}\ cm^{-3}$, $T_e \sim 2.5$ eV $\nu_{en} \sim 17$ MHz, $f_p \sim 1.3$ GHz and $f_c \sim 0.63$ GHz. The wave launcher (Fig. 1(b)) is a 50 Ω parallel plate transmission line tapered to match into a coaxial line with better than -30 dB reflection loss. The plates are separated by about 1 cm. Hence, when a 3 KV pulse is applied to the coaxial line with



(a) EXPERIMENTAL SETUP



(b) PARALLEL PLATE TRANSMISSION LINE

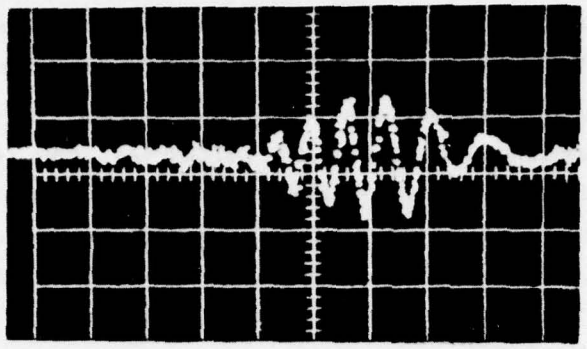
Fig. 1.

about 4dB loss, a 1.9kV/cm electric field would appear between the plates. A hole is bored in each plate and the plasma tube of radius 0.66cm with the conducting waveguide is inserted perpendicular to the plates to allow the applied impulsive field in the axial direction shown in Fig. 1(b). The polarity of the field cannot be reversed because of the ground loop requirements of equipment used in the setup. To insure tight coupling at the input, a pair of copper rings are built into the discharge tube so that direct contact of plasma and the plates is achieved. The electric field E_z of the wavepacket propagating along the column is picked up by a similar structure and is observed on a sampling scope (Textronix 564). All experimental phenomena were observed during a period less than 50ns which is short compared with the fastest collision time of the plasma (> 100ns). Thus, the observations are free from any undesirable effects such as ionization and heating usually accompanying high amplitude non-linear experiments. A series of wide band attenuators are inserted in both the input and output lines to allow

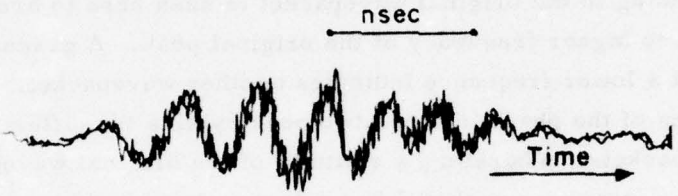
a wide range of input field strengths to be used. Each time an attenuator is removed from the input line, it is added to the output line to maintain the overall loop gain constant. The plasma background density is monitored continuously by a microwave cavity for observing long term stability. The short term stability of the system is extremely good when no axial magnetic field is applied. This is demonstrated by the oscillograph of Fig. 2(a) which is one typical observed wavepacket at linear level. Each dot on the oscillogram represents one real time experiment. These oscillograms can often be repeated many times without noticeable changes. From the time required to complete one oscillogram (several seconds) and the repetition rate of the BBG (~100 pps), one can estimate that the experiments are repeatable thousands of times in real time. When a DC axial B field is applied, some small striations (~100 Hz) will occur. Therefore care must be exercised in processing the data. To improve accuracy, an X-Y recorder driven by the sampling scope was used. The output waveforms were recorded on 14" x 10" recording papers. In a slowly recorded waveform the striations will show up as slight up and down jitter of the recording pen as shown typically in Figure 2(b). Because of the large display, there is no difficulty to retrace the waveforms to smooth out the jitters. The smoothed waveforms were digitalized for computer Fourier analysis with the aid of a carefully devised system which incorporates a calibrated grid, a set of high quality helipots and a digital voltmeter. The resulting amplitude data are accurate to the second digit with the third digit approximate. The time base is as accurate as that of the sampling scope. We note that if the jitter were included in the data, it would only appear as a fictitious very high frequency component in the spectra.

B. Results

Figure 3 shows a typical sequence of the smoothed out output waveforms taken at a distance of 65 cm from the input, corresponding to the plasma background given earlier. Other conditions give minor variations from this sequence. Figure 3(a) shows the response at the linear level, which is the familiar Airy function[3,6] like wavepacket. As the input impulse field is increased, gross non-linear phenomena begin to occur. From Figs. 3(b)-(d), one observes that signals arrive sooner than the linear waves and considerable distortion of the original wavepacket begins to take place. Figure 3(d) indicates an interference of waves. To analyze these waveforms, computer Fourier transforms of the signals were performed. At least 10 data points were used in one period of oscillation. The spectral amplitude and phase for each waveform in Fig. 3 are shown in Figure 4. At linear level Fig. 4(a), only one peak (C) is shown corresponding to the wavepacket of Fig. 3(a). The slope of the phase curve near the peak represents a time delay from some reference point. The slight curvature of the phase characteristic indicates a slight amount of FM present in the wavepacket. Using the



(a)



(b)

Fig. 2.

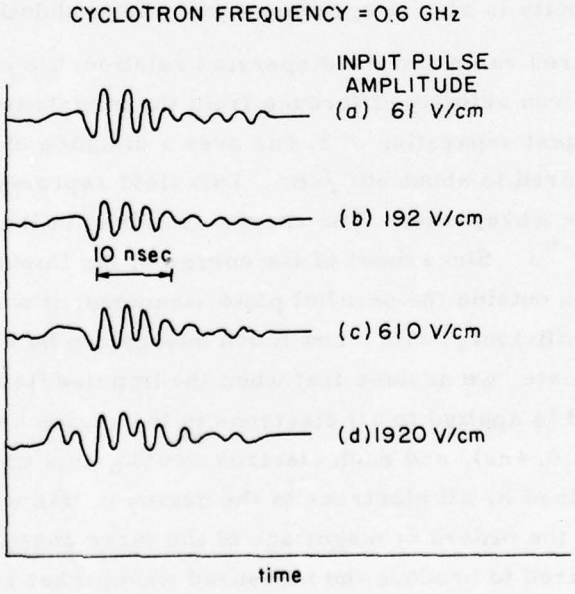


Fig. 3. Typical sequence, distance: 65 cm.

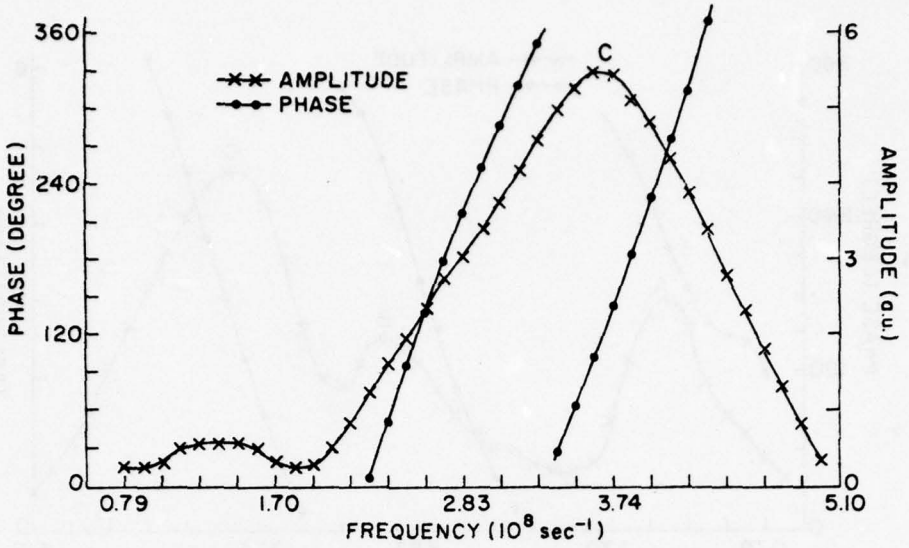
calculated linear group velocity and the phase slope, this reference point can be established and can be used for determining the group velocities of the other wavepackets.

At higher amplitude two distinct phenomena occurred (Figures 4(b)-(d)). First is the appearance of a low frequency peak at f = .147 GHz(A). The amplitude of this peak increases more rapidly as the input level is increased (recall that the loop gain is kept at a constant). There is a slight low shift of frequency as amplitude is increased. From the phase slope (i.e., time delay), we can show a linear amplitude dependence of velocity. No further analysis of this packet was made but we strongly suspect that it belongs to the class of solitary waves.

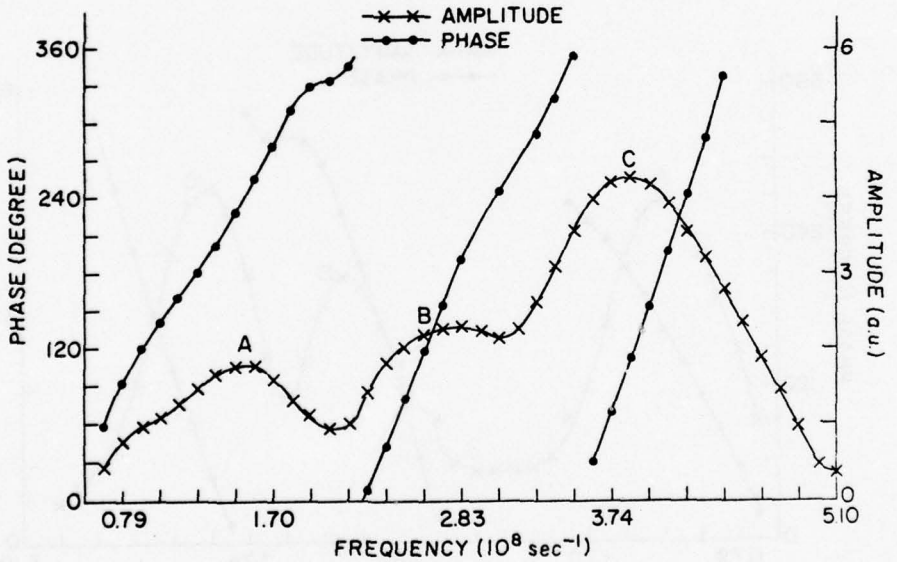
The main subject of this report is on the change associated with the peak C. The peak C corresponding to the original wavepacket is seen here to break into two. There is a definite shift to higher frequency of the original peak. A gradual appearance of another peak B at a lower frequency indicates another wavepacket. The difference between the slopes of the phase of these two peaks yields the difference in the times of arrival of these packets. Therefore a splitting of the original wavepacket has taken place. The higher frequency packet C has a longer delay and hence a lower velocity.

According to the modulation theory,[1,2] the difference between the two group velocities is proportional to the amplitude as shown by Equation (3). Hence the difference between the arrival times Δt should also be proportional to the amplitude to the first order. This is shown in Figure 5. The observation that the higher frequency wavepacket travels at a lower velocity is also in agreement with the modulation theory.

From the measured value and the dispersion relation, we can calculate the required field amplitude for a given velocity difference from the modulation theory. It is shown that to achieve the largest separation of 3.4 ns over a distance of 65 cm, the amplitude of the wavepacket required is about 50 C/cm. This field represents a total energy of about 1.3×10^{-9} J in the wavepacket. The energy contained in the exciting electric field impulse is about 3×10^{-5} J. Since most of the energy of the impulse goes into the dummy load and radiates to outside the parallel plate structure, it would be of interest to estimate the coupling efficiency, i.e., how much energy can be coupled into the plasma. As a rough bound estimate, we assume that when the impulse field arrives at the plasma tube, the 2 kV/cm field is applied to all electrons in the region between the plates during the pulse duration ($\tau \approx 0.4$ ns), and each electron would gain a momentum of $eE\tau$. The total kinetic energy gained by all electrons in the region in this way is of the order of 9×10^{-7} J. Comparing the orders of magnitude of the three energy levels, we feel that the 50 V/cm field required to produce the measured wavepacket separation is certainly reasonable for the experimental setup used.

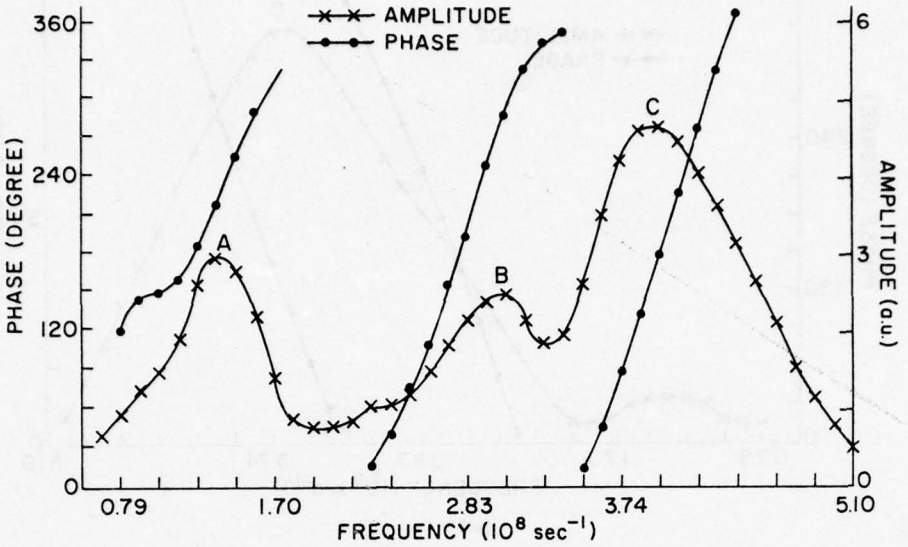


(a)

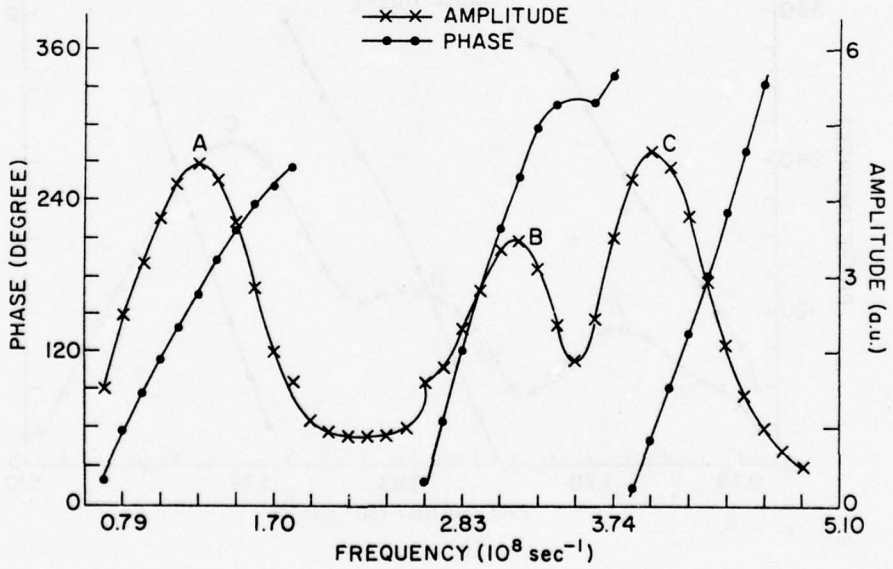


(b)

Fig. 4. Fourier spectra of Figure 3.



(c)



(d)

Fig. 4. Fourier spectra of Fig. 3 (Continued).

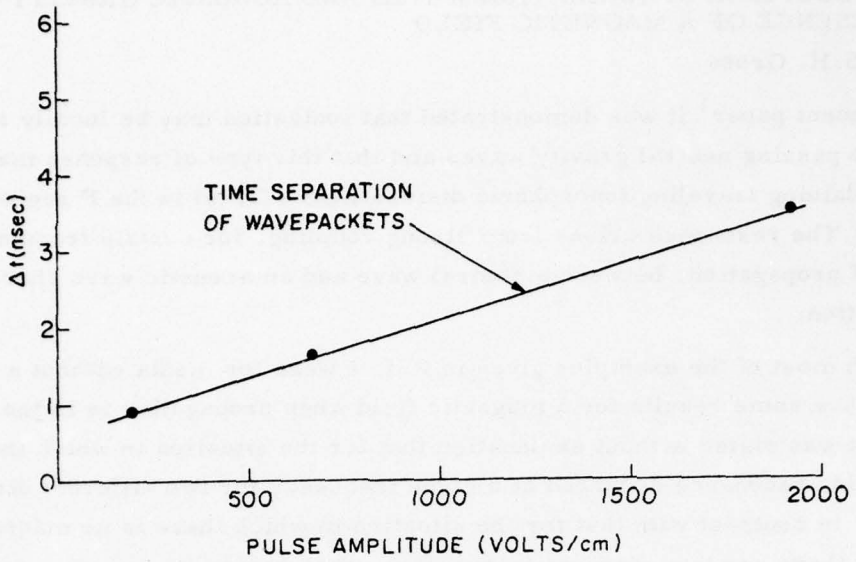


Fig. 5. Time separation vs. amplitude.

In conclusion, we believe that we have shown sufficient evidence that the wave-packet splitting effect has been observed as predicted by the modulation theory. This new experimental observation should complement other efforts such as computer simulations to offer some confidence in the application of the modulation theory to other nonlinear wave problems.

Air Force Office of Scientific Research
AFOSR-74-2668

B.R. Cheo

REFERENCES

1. G.B. Whitham, "Linear and Nonlinear Waves," 519-520 (New York: John Wiley and Sons, 1974).
2. G.B. Whitham, Ibid. 489-491.
3. E. Kunhardt, "An Experimental and Theoretical Study of the Propagation of High Amplitude Pulses in a Bounded Magneto Plasma," Ph.D. Dissertation, Polytech. Inst. of New York (June 1976).
4. A. Trivelpiece and R.W. Gould, J. Appl. Phys., 30, 1784-1793 (1959).
5. P. Pleshko and I. Palocz, Phys. Rev. Lett., 22, 1201-1204 (1969).
6. D.L. Landt, et al., Plasma Physics, 16, 407-412 (1974).

RESONANT COUPLING OF IONIZATION WAVES AND ACOUSTIC GRAVITY WAVES IN THE PRESENCE OF A MAGNETIC FIELD

H. Eun and S.H. Gross

In a recent paper[1] it was demonstrated that ionization may be locally resonant at each level to passing neutral gravity waves and that this type of response may be significant in explaining traveling ionospheric disturbances (TID's) in the F region of the ionosphere. The resonance arises from strong coupling, for certain frequencies and directions of propagation, between a neutral wave and an acoustic wave characteristic of the ionization.

Though most of the examples given in Ref. 1 were for media without a magnetic field, they gave some results for a magnetic field when propagation is in the magnetic meridian. It was stated without explanation that for the situation in which there is a magnetic field, resonance occurred at a given frequency for two different directions of propagation, in contrast with that for the situation in which there is no magnetic field and in which there are two resonant frequencies, each having its own direction. In this contribution we provide fuller details for a magnetic field. How two resonant directions arise is explained graphically. In addition, results for propagation out of the meridian plane are presented which show that the general features of meridional plane propagation are mostly preserved, though the direction is at an angle to that plane.

In an early paper[2] a similar problem was treated. The analysis was incorrect because it neglected the effects of ions, as pointed out in Reference 3. Two figures in Ref. 2, however, on examination show the possibility of two resonant directions, though the presentation is somewhat different from that given here. Because we believe the possible existence of two resonant directions is more important than may have been appreciated, we emphasize this feature in our treatment and discussion.

Though still other authors more recently have analyzed the passage of acoustic gravity waves through the ionosphere,[4, 5] the resonant coupling phenomenon was not recognized. Reference 6 also reviewed and discussed original work on resonant coupling but did not develop the theory further. The relevance of these previous efforts was discussed in Ref. 1 and will not be repeated here.

As was discussed in the work reported in Ref. 1, some amount of losses may actually enhance coupling, though the resulting waves are damped on propagating away from the coupling region. An example showing resonant coupling with thermal conductivity losses was presented in that paper which demonstrated that resonant coupling is unaffected by losses. The role of viscosity in the F region was also considered and deemed unimportant in relation to thermal conductivity as a loss mechanism. In view of these previous results, viscosity and thermal conductivity were neglected for simplicity in the analysis and computations presented here.

With the exception of assumptions related to the magnetic field the method of analysis is that described in Reference 1. As before, we linearize by a perturbation analysis, use Cartesian geometry, assume thermodynamic equilibrium, neglect ionization production and loss, assume time stationary background quantities, treat the ionization as a single fluid of half the ion mass with a content minor in comparison with the neutrals, and utilize the plane wave approximation. Though the last assumption is simpler to use, the same essential results are obtainable from a WKB analysis similar to that given in Reference 7. Additional assumptions associated with the magnetic field are that the field lines are straight but inclined as given by the dip angle I and that motion transverse to the field lines is negligible. In the F region the electron and ion collision frequencies are taken to be much smaller than their respective gyrofrequencies. These various assumptions, however, are such that hydromagnetic waves are eliminated from consideration, a matter for further examination.

Examples are given showing two resonant directions. It is also shown that a strong coupled response of ionization is possible over a very large frequency band and for a fair range of directions that in some cases may be quite wide. The possibility is illustrated that a peaked response may occur over an altitude range of the order of 100 km for a wave of a particular frequency and direction. Such an altitude range would suggest that strong coupling is not as localized as had originally been thought. Though this suggestion must be checked by a more rigorous analysis, its correctness would considerably augment the importance of the resonance mechanism in explaining observed perturbations.

First, an analysis is presented in which the ionization density perturbation is derived. The resonance and its properties are then treated. Calculations for typical cases are presented afterward to show the effects of varying both the dip angle and the angle between the plane of propagation and the meridian plane.

A. Analysis

A Cartesian coordinate system is used as shown in Fig. 1, the x and y axes being directed horizontally toward magnetic north and west, respectively, and the z axis directed vertically upward. The magnetic flux density vector $\underline{B}_0$ is taken to be lying in the xz plane with a dip angle I that is positive in the northern hemisphere. The perturbing neutral wave propagates in the direction of its wave number vector $\underline{k}$ located with respect to the coordinate axes by the colatitudinal angle θ measured from the z axis and by the azimuthal angle ϕ measured from the x axis. Quantities having the subscripts 0 and 1 represent background and perturbation quantities, respectively. The subscripts n, e, and i designate neutral, electron, and ion fluid quantities, respectively. A quantity without these subscripts refers to the entire medium. The symbols p, ρ, n, T,

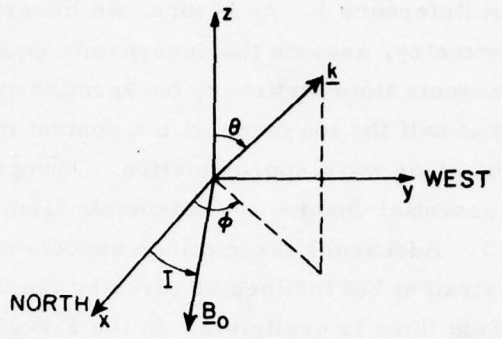


Fig. 1. Coordinate system.

$\underline{C}$, m, g, and H represent in that order pressure, mass density, number density, temperature, peculiar or diffusion velocity, mass, gravitational acceleration, and scale height. Peculiar velocity $\underline{C}$ is specified in relation to the mean mass velocity $\underline{c}_1$, which is taken to be the particle motion caused by the perturbing neutral acoustic gravity wave. D_a is the ambipolar diffusion coefficient, and ν_{ij} is the momentum transfer collision frequency of species i with species j.

A two-fluid system is used, one fluid being for the ionization and the other for the neutrals. The equations for the neutrals may be replaced by those for the entire medium, since the ionization content is very small in relation to that of the neutrals. The background ionization is assumed to undergo ambipolar diffusion along the field lines with peculiar velocity

$$\underline{C}_{i0} = -D_{a0}\left(\frac{1}{p_{i0}} \nabla_b p_{i0} - \frac{\sin I}{2 H_n}\right) \underline{a}_b \tag{1}$$

where ∇_b is the scalar component of the del operator along the field line, $\underline{a}_b$ is the unit vector along that line, and $\underline{C}_{i0}$ is the background ambipolar diffusion velocity. The background neutral motion, however, is taken to be zero, and the unperturbed neutrals are distributed hydrostatically. Neutral particles move only with the wave perturbation velocity $\underline{c}_1$ which is taken to be the mean mass velocity.

The perturbed motion of the ionization is described by the following set of hydrodynamic equations:

$$\frac{\partial}{\partial t} W_{i1} + \nu_0 W_{i1} + 2\left(\cos I \frac{\partial}{\partial x} - \sin I \frac{\partial}{\partial z}\right) p_{i1} - \rho_{i1} g \sin I$$

$$= -c_{1b} \nabla_b W_{i0} - W_{i0} \nabla_b c_{1b} - \rho_{i0} \frac{\partial c_{1b}}{\partial t} - W_{i0} \nabla_b c_{1b} - \nu_1 W_{i0} \tag{2}$$

$$\frac{\partial}{\partial t}\rho_{i1} + \nabla_b W_{i1} + \nabla_b(\rho_{i0} c_{1b}) = 0 \tag{3}$$

$$p_{i1} = n_{i1} K_b T_0 + n_{i0} K_b T_1 \tag{4}$$

Here K_b is Boltzmann's constant, $W_i = \rho_i C_i$ is the ionization momentum flux along the field line, and c_{1b} is the component of the neutral particle velocity $\underline{c}_1$ along the field line. Equation (2) is the momentum equation for the component along the field line, Eq. (3) is the continuity equation, and Eq. (4) is the perturbed equation of state. The temperature perturbation is determined only by the neutral medium's response to the acoustic gravity wave because of the assumption of thermodynamic equilibrium. Wave quantities for the neutral medium obey the well-known acoustic gravity wave dispersion relationship, and perturbation quantities that are solely due to neutral waves in Eqs. (2), (3), and (4) are regarded either as source terms driving the ionization or as terms coupling the neutral motion to the ionization. The motion of the neutrals is assumed to be known, and the purpose of the analysis is to find the response of the ionization to this perturbation.

Since we seek a plane wave solution, we take $(\partial/\partial x) = ik_x$ and $(\partial/\partial z) = ik_z$. Then, with $(\partial/\partial t) = -i\omega$, Eq. (2) through Eq. (4) may be utilized to yield the ionization density perturbation as given by

$$\frac{n_{i1}}{n_{i0}} = \left[\frac{1}{n_{i0}}\nabla_b(n_{i0}c_{1b}) + M_1^B + M_2^B\right]\left\{i\omega + D_{ao}^* \cdot \left[-ik_b\frac{\sin I}{2H_n} - k_b^2 - 2ik_b\frac{T_0'}{T_0}\sin I + \frac{T_0''}{T_0}\sin^2 I + \frac{\nu_0'}{\Omega_0}\sin I\left(ik_b - \frac{T_0'}{T_0}\sin I - \frac{\sin I}{2H_n}\right)\right]\right\} \tag{5}$$

where

$$M_1^B = -\frac{1}{n_{i0}}\nabla_b\left(\frac{\nu_1}{\Omega_0} n_{i0} C_{i0}\right) - D_{a0}^*\frac{T_1}{T_0}\left[\frac{n_{i0}''}{n_{i0}}\sin^2 I - 2k_b\frac{n_{i0}'}{n_{i0}}\sin I - k_b^2 + \frac{\nu_0'}{\Omega_0}\sin I\left(ik_b - \frac{n_{i0}'}{n_{i0}}\sin I\right)\right] \tag{6}$$

$$M_2^B = -\frac{1}{n_{i0}}\nabla_b\left\{\frac{1}{\Omega_0}\left[(c_{1b}\nabla_b)n_{i0}C_{i0} + (\nabla_b c_{1b})n_{i0}C_{i0} - i\omega n_{i0}c_{1b} + (n_{i0}C_{i0}\nabla_b)c_{1b}\right]\right\} \tag{7}$$

$$k_b = k_x \cos I - k_z \sin I \tag{8}$$

$$\nu_0 = \nu_{in0} + \beta \nu_{en0}, \quad \beta = \frac{m_e}{m_i} << 1, \quad \Omega_0 = -i\omega + \nu_0, \quad D^*_{a0} = \frac{2K_b T_0}{m_i \Omega_0}$$

a modified ambipolar diffusion coefficient, and $\nu_1 = \nu_{in1} + \beta \nu_{en1}$. The prime and double prime notations above background quantities mean first and second derivatives with respect to z, respectively. Equation (5) was given in Ref. 1 without derivation.

Equation (5) illustrates the dependence of the ionization density perturbation on background quantities as well as the neutral wave perturbation quantities $\underline{c}_1$ and T_1. The neutral density perturbation also enters the formulation through ν_1.

B. Resonance

Resonance occurs when the denominator in Eq. (5) is zero. It should be noted that the dispersion relationship for ionization acoustic waves with particle motion constrained along the field line is given by the same expression when it is set equal to zero.

The characteristics of resonance are most easily demonstrated for the case of a lossless isothermal atmosphere when we take $\omega << \nu_0$. Eigenvalues of the ionization acoustic wave may then be obtained from

$$i\omega + D_{a0}\left[-ik_b \frac{\sin I}{2H_n} - k_b^2 + \frac{\nu_0'}{\nu_0} \sin I\left(ik_b - \frac{\sin I}{2H_n}\right)\right] = 0 \tag{9}$$

which is obtained from the denominator of Eq. (5) by using these simplifying assumptions.

The $\underline{k}$ vector of the neutral wave is complex, since the z component has an imaginary part $(-1/2H_n)$ that results in the growth of the relative perturbation density and the perturbation velocity with altitude. On separating Eq. (9) into real and imaginary parts, one obtains

$$k_{br}^2 = \frac{3}{2H_n^2} \sin^2 I \tag{10}$$

$$\omega = \frac{5}{2} \frac{D_{a0}}{H_n} \sin I \, k_{br} \tag{11}$$

where k_{br} is the real part of k_b. The ω in Eq. (11) will be the critical coupling or resonance frequency ω_c for the given dip angle and altitude if the acoustic gravity wave number vector $\underline{k}$, with a component along the magnetic field k_b, and ω_c satisfy the acoustic gravity wave dispersion relationship. Unlike the similar situation without a magnetic field, as given in Ref. 1, k_x does not enter the ionization relationships Eqs.

(10) and (11) obtained from Eq. (9). There is now only one wave vector component, k_{br}. It is determined entirely by the dip angle, neutral particle scale height (or temperature), and the collision frequency through the diffusion coefficient D_{a0}. The frequency $\omega = \omega_c$ in Eq. (11) also depends on these same quantities. In the northern hemisphere, I is taken as positive, though the field lines point downward. From Eq. (11) it is obvious that k_{br} must be positive; that is, the real part of the component of $\underline{k}$ along the field line must be downward along the field line. This would imply that the real part of the vector $\underline{k}$, namely $\underline{k}_r$, would very likely be downward as well. In the southern hemisphere, sin I is negative and k_{br} must be negative or downward, too. The neutral wave vector $\underline{k}$ or $\underline{k}_r$ is also likely to be downward.

The critical values of $\underline{k}$, k_z, and k_x are found for a given altitude and dip angle by simultaneously solving Eqs. (10) and (11) with the dispersion relationship for acoustic-gravity waves.[1] This procedure yields, in general, two directions for resonance for the same frequency $\omega = \omega_c$, given by Equations (10) and (11). One may proceed to treat this problem analytically, but it is preferable for the purposes here to utilize geometrical diagrams to illustrate the nature of the resonance and how it changes with dip angle. It is best to show these relationships for propagation in the plane of the magnetic meridian. The characteristics are essentially the same for propagation out of the meridian, as one may show by projecting the wave vector surface for the neutral medium into the meridian plane and proceeding as described below.

The relationship between $\underline{k}$, $\underline{k}_r$, and k_{br} must be appreciated. Here $\underline{k}_r$, which is the real part of $\underline{k}$, is given by $\underline{k}_r = \underline{k} - ik_{zi}\underline{a}_z$, where k_{zi} is the imaginary part of k_z and $\underline{a}_z$ is the unit vector along z. Thus $\underline{k}_r = \underline{a}_x k_x + \underline{a}_z k_{zr}$ where k_{zr} is the real part of k_z and $\underline{a}_x$ is the unit vector along x. Then k_{br}, the real part of k_b, is the component of $\underline{k}_r$ along the magnetic field line. We may then take the vector $\underline{k}_{br}$ to be the vector along the direction of $\underline{B}_0$ with magnitude k_{br}. Given $\underline{k}_{br}$, all possible $\underline{k}_r$ with this component k_{br} must have their vector arrowheads lying along a line perpendicular to $\underline{k}_{br}$, the line passing through the arrowhead of $\underline{k}_{br}$. The construction for the northern hemisphere is shown as line L in Figure 2.

Figure 2 represents the $\underline{k}_r$ vector space in the meridian plane. Also shown in the figure are two hyperbolic curves representing the gravity wave dispersion relationship in this plane for the frequency $\omega = \omega_c$ and the altitude of interest. Line L intersects each hyperbola as shown, and these intersections give the wave directions $\underline{k}_{rc}^{(1)}$ and $\underline{k}_{rc}^{(2)}$ that satisfy both Eqs. (10) and (11) and the gravity wave dispersion relationship. Any wave of frequency $\omega = \omega_c$ traveling in the neutral medium in either of these two directions at the particular altitude would result in resonant coupling to ionization acoustic waves. These acoustic waves would then leave the coupling region and propagate on their own.

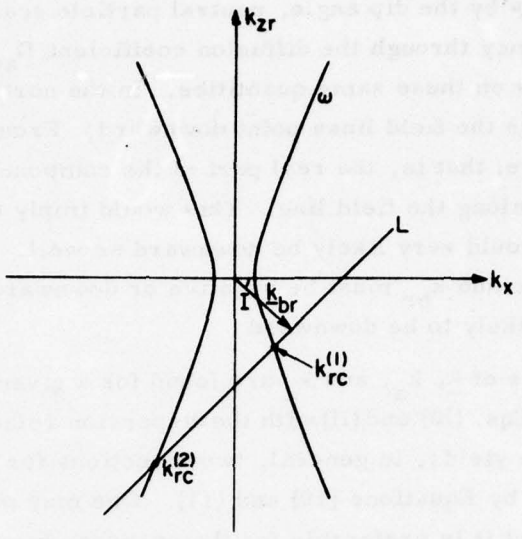


Fig. 2. Existence of double resonances for a given dip angle for the gravity wave branch.

The geometrical analysis also exhibits the changes in the nature of the resonance as the dip angle is increased from small values to larger values for a given altitude. The frequency $\omega = \omega_c$ increases with $\sin^2 I$ and decreases inversely with background density (Equations (10) and (11)). For small values of I, ω_c is relatively small at any given altitude, and the required waves in the neutral medium are likely to be in the gravity wave range. The hyperbolas are steep, and intersections, or wave coupling directions, are in the third and fourth quadrants of the $\underline{k}_r$ plane for the northern hemisphere, that is, in the downward direction to the north (+x or $+k_x$) and south (-x or $-k_x$).

It is apparent from Fig. 2 that the slope of line L for small I is smaller than the asymptotic slope of the first-quadrant branch of the hyperbola. It can be shown that the latter slope decreases more rapidly as I increases than the slope of line L. At some value of I the slopes are the same and there is only one intersection which is in the fourth quadrant. Further increases in I produce intersections in the first and fourth quadrant, as shown by L_1 in Fig. 3 for frequency ω_1, yielding both upward and downward northerly directions for resonance. As I increases further, the hyperbolas move toward higher values of k_x, and eventually, line L is just tangent to the hyperbola, as shown by L_2 at frequency $\omega_2 > \omega_1$ in Figure 3. Resonance is not possible in the gravity wave branch for further increases in I, so that L_2 represents the upper cutoff of the phenomenon in this branch. The cutoff frequency is less than the Brunt-Väisälä frequency at the particular altitude. At high enough altitudes, further increases in I may permit resonance at

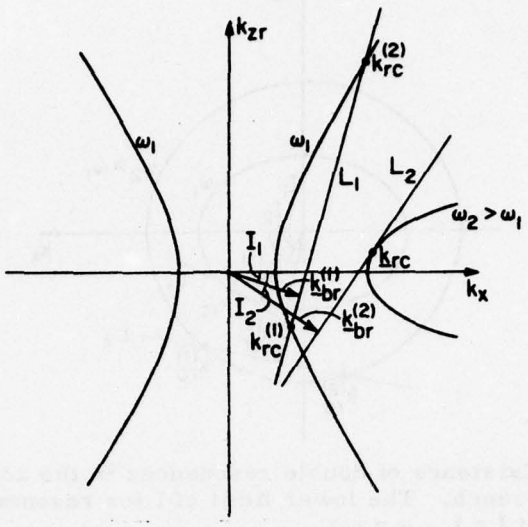


Fig. 3. Same as Fig. 2 but for larger values of dip angle. The upper limit of I for resonance in the gravity wave branch is shown at $I = I_2$, $\omega = \omega_2$.

frequencies in the acoustic branch of the acoustic-gravity wave dispersion relationship, as shown in Fig. 4, in which line L intersects ellipselike wave vector curves that are characteristic of this branch. L_1 for $\omega = \omega_1$ in Fig. 4 represents the low-frequency cutoff resonance in the acoustic branch. Here ω_1 is greater than the acoustic branch cutoff frequency ω_a. L_2 exhibits two resonant directions for frequency $\omega_2 > \omega_1$. Acoustic branch coupling may possibly occur for altitudes above 500 km. It should be noted that Ref. 8 reports localized ionospheric disturbances at acoustic frequencies. Acoustic branch coupling within the assumptions here is conceivable up to $I = 90^\circ$.

C. Calculations

Calculations are made for various altitudes and various directions of propagation of the acoustic-gravity wave in the neutral medium. The background model that is utilized is the same as that given in Figs. 1 and 2 of Reference 1.

Figures 5 and 6 are polar plots of the ionization perturbation, normalized to the zero frequency response, in the meridional plane in the northern hemisphere for an altitude of 400 km in Fig. 5 and 250 km in Figure 6. A normalized $|n_{i1}/n_{i0}|$ is plotted in each figure versus θ, the colatitudinal angle in Figure 1. Zero degrees is along the upward vertical. The angles are labeled positively from 0° to 180° for the right-hand or north-directed (+x) side and negatively from 0° to -180° for the left-hand or south-

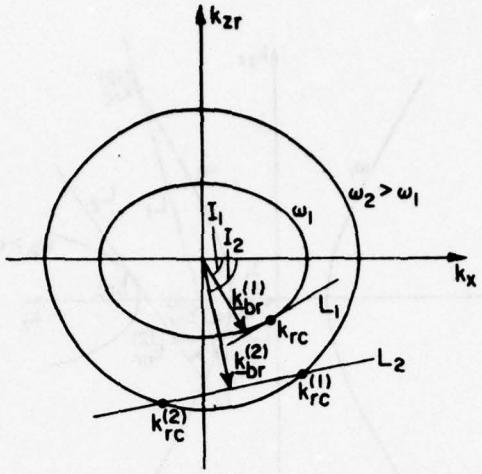


Fig. 4. Existence of double resonances in the acoustic wave branch. The lower limit of I for resonance is shown at $I = I_1$, $\omega = \omega_1$.

directed (-x) side. North is along $\theta = 90^\circ$, and south along -90°. The dip angle for Figs. 5 and 6 is 40°.

One may observe in each figure that there are six curves, which may be divided into two sets of three curves each, that are placed nearly symmetrically with respect to each other about the vertical axis. One curve of each set is a resonance curve having a theoretically infinite peak reaching the edge of the graph in the figure. The resonance frequency ω_c is given in the caption for each figure. The other two curves in a set are for frequencies that are 50% greater and 50% less than the resonance frequency. Though infinite responses are not obtained for these frequencies, which are well off resonance, they still exhibit large peaks (the scale is logarithmic). The directions of their peaks, however, are not the same as the direction for resonance. All are downward pointing, but the direction of the peak for the 50% greater frequency is the furthest from the vertical of the three frequencies, whereas that for the 50% lower frequency is closest to the vertical.

Peaked responses within each set are evident in Fig. 5 with a 15° spread of direction for a 3:1 frequency range. The two sets, or the two resonant directions, combined account for a 30° band of directions. The 3:1 frequency range was arbitrarily chosen, and one may expect strong ionization response over a greater angular range for a larger frequency band. Responses for the same altitude for larger dip angles yield much wider angular spreads than the angular spread in Figure 5. Peaked responses within each set as described for Fig. 5 are found to be as great as 50° about the resonant

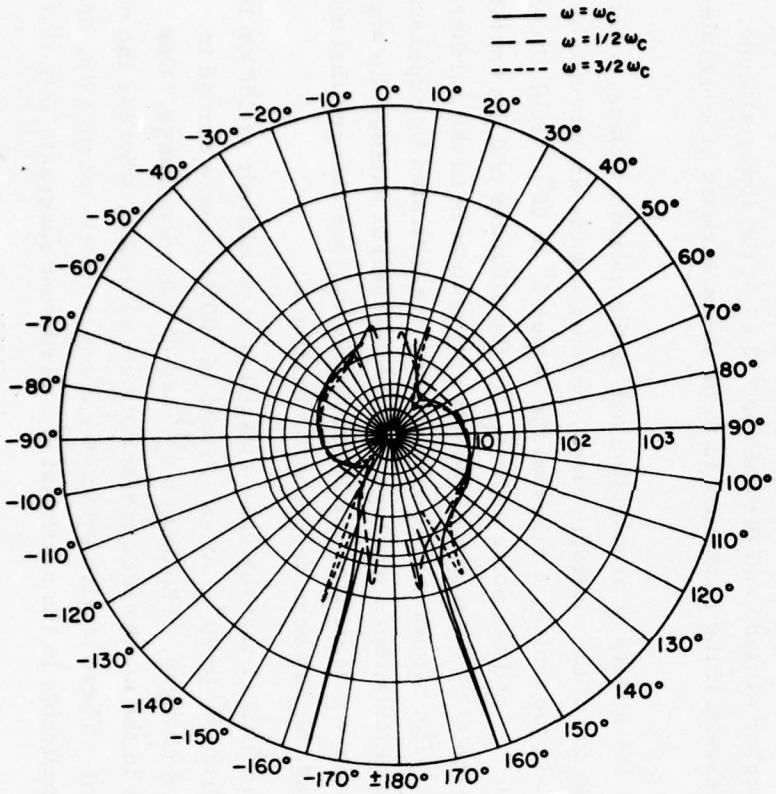


Fig. 5. Normalized polar plot of $|n_{i1}/n_{i0}|$ in the xz plane for three frequencies, for an altitude of 400 km and dip angle 40°. Values are normalized relative to the value at $\omega = 0$. The resonant frequency ω_c is 1.79×10^{-3} rad/sec, and its directions are $\theta_c = 162.60°$ for northward propagation and $\theta_c = -164.10°$ for southward propagation.

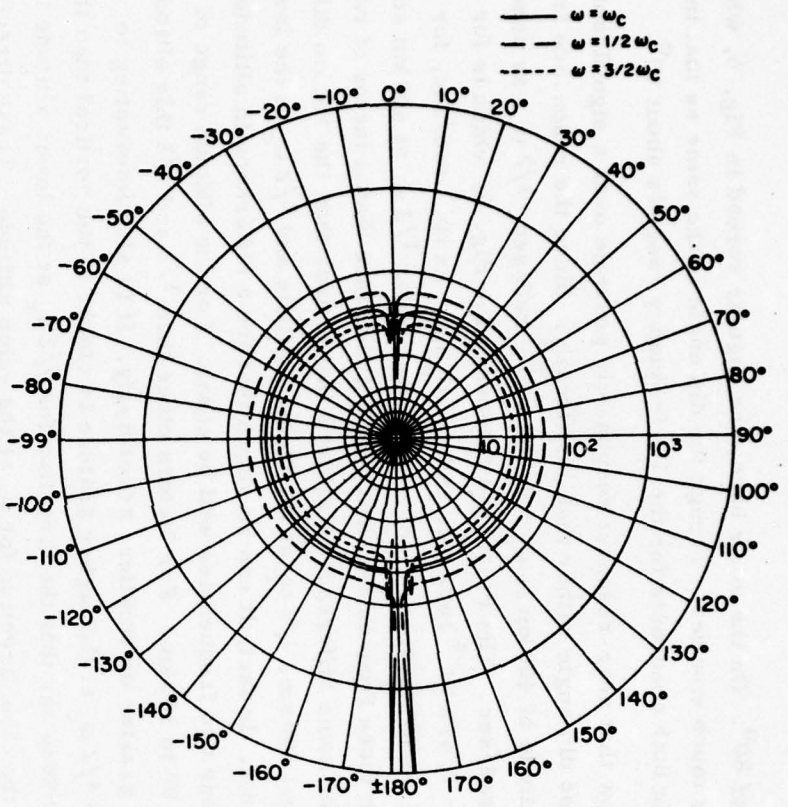


Fig. 6. Same as Fig. 5 but for an altitude of 250 km $\omega_c = 2.47 \times 10^{-4}$ rad/sec. $\theta_c = -178.09°$ for northward propagation and $\theta_c = -178.11°$ for southward propagation.

direction for a dip angle of 80°. On the other hand, the angular spread in Fig. 6, which is for a lower altitude, is much smaller. Though the dip angle is the same as that in Fig. 5, the total spread in both quadrants for the 3:1 frequency range is about 10°.

Figures 5 and 6 show that near-resonant response is possible over a significant altitude range for the same dip angle rather than just locally. Since the resonance frequency of Fig. 5 at an altitude of 400 km is $\omega_c = 1.794 \times 10^{-3}$ rad/sec, $1/2\ \omega_c$, as used in Fig. 5, is 8.97×10^{-4} rad/sec. The resonance frequency for Fig. 6, which is for an altitude of 250 km, is $\omega_c = 2.47 \times 10^{-4}$ rad/sec, and $3/2\ \omega_c$ is 3.7×10^{-4} rad/sec, for Figure 6. Strong peaking occurs at these frequencies, that is, at $1/2\ \omega_c$ at 400 km and $3/2\ \omega_c$ at 250 km. Though these frequencies differ by a little more than a factor of two, it is quite apparent that for some difference in altitude that is less than the 150 km difference between 250 km and 400 km, $1/2\ \omega_c$ at the higher altitude and $3/2\ \omega_c$ at the lower altitude may indeed be equal. In fact, it can be shown that for a difference in altitude equal to about $1.1\ H_n$, these two frequencies will be equal. For the altitude range of interest, $1.1\ H_n$ is some 90 to 100 km. For factors other than $1/2$ and $3/2$ this altitude range would be somewhat greater or smaller, accordingly. It is also interesting to note that the direction for $1/2\ \omega_c$ at the higher altitude is closer to the vertical than the direction for ω_c at that altitude and that the direction for $3/2\ \omega_c$ at the lower altitude is further from the vertical than the direction for ω_c at the same altitude. These directions, the one for $1/2\ \omega_c$ at the higher altitude and the one for $3/2\ \omega_c$ at the lower altitude, tend to overlap, as is necessary if the resonance is to be characteristic of a sizable range of altitude.

Tables I and II show the effects of propagation out of the meridian plane, ϕ being used to represent the angle out of the meridian plane. Propagation is assumed to be in the northern hemisphere. In the tables, ϕ takes on the values 0° to 90°. Table I compares resonance values for k_{rc} and the polar angle θ_c for the altitudes and dip angles of Figures 5 and 6. The sign of θ_c determines whether propagation is in the northerly or the southerly direction. The two wave vector directions are indicated by superscripts (1) and (2). Table II gives calculations for a high altitude (600 km) at various dip angles. This altitude is chosen because calculations for it exhibit the various changes that may occur as dip angle is varied, as has been described in Section B.

It may be observed in Table I that the characteristics at an altitude of 250 km barely change as ϕ increases from 0° to 90°. At an altitude of 400 km, some change in $k_{rc}^{(1)}$ and $k_{rc}^{(2)}$ is evident as ϕ changes. The angles $\theta_c^{(1)}$ and $\theta_c^{(2)}$ hardly change. One direction of propagation is in the northwest quadrant at both altitudes, whereas the other is in the southeast quadrant. There is also symmetry about the north-south axis, and so the same results are applicable in the northeast and southwest quadrants with their corresponding values of ϕ.

TABLE I. Wave number k_{rc} and angle θ_c versus ϕ for two altitudes and for I = 40°.

ϕ	0°	10°	20°	40°	60°	80°	90°
			250 km, $\omega_c = 2.47\times10^{-4}$				
$k_{rc}^{(1)}$	2.102×10^{-5}	2.103×10^{-5}	2.107×10^{-5}	2.121×10^{-5}	2.143×10^{-5}	2.170×10^{-5}	2.185×10^{-5}
$\theta_c^{(1)}$	178.09	178.09	178.09	178.09	178.10	178.10	178.10
$k_{rc}^{(2)}$	2.275×10^{-5}	2.273×10^{-5}	2.269×10^{-5}	2.253×10^{-5}	2.229×10^{-5}	2.200×10^{-5}	2.185×10^{-5}
$\theta_c^{(2)}$	-178.11	-178.11	-178.11	-178.11	-178.11	-178.10	-178.10
			400 km, $\omega_c = 1.79\times10^{-3}$				
$k_{rc}^{(1)}$	1.223×10^{-5}	1.229×10^{-5}	1.245×10^{-5}	1.309×10^{-5}	1.419×10^{-5}	1.576×10^{-5}	1.672×10^{-5}
$\theta_c^{(1)}$	162.60	162.61	162.66	162.84	163.09	163.36	163.50
$k_{rc}^{(2)}$	2.524×10^{-5}	2.505×10^{-5}	2.450×10^{-5}	2.260×10^{-5}	2.018×10^{-5}	1.779×10^{-5}	1.672×10^{-5}
$\theta_c^{(2)}$	-164.10	-164.10	-164.07	-163.98	-163.83	-163.62	-163.50

In Table II, one sees at I = 7° that θ_c barely changes value. Both $\theta_c^{(1)}$ and $\theta_c^{(2)}$ correspond to steep third- and fourth-quadrant directions. Both values for k_{rc}, on the other hand, change by factors between 2 and 3 as ϕ increases from 0° to 90°. At I = 10°, one wave, as given by $k_{rc}^{(1)}$ and $\theta_c^{(1)}$, hardly changes direction, but the wave number increases by a factor of more than 2 as ϕ increases. The direction of propagation is in the fourth quadrant. The other wave, for small ϕ, is in the first quadrant. However, when ϕ = 20°, we find that it has turned steeply downward in the third quadrant. This means that for some angle between ϕ = 10° and ϕ = 20° the slope of line L in Fig. 2 must equal the asymptotic slope of the projected hyperbola in the meridian plane. For this case there is only one resonance and it is in the fourth quadrant. As ϕ increases further, the angle of propagation barely changes, though $k_{rc}^{(2)}$ decreases. For I = 20° the same features as those for I = 10° are observed. However, $\theta_c^{(1)}$ increases from about 114° to 135°, the wave becoming steeper with increasing ϕ. The angle $\theta_c^{(2)}$ lies in the first quadrant out to at least ϕ = 60°. It changes to the third quadrant between ϕ = 60° and ϕ = 80°. For I = 40° the frequency ω_c has increased to a value that must lie in the acoustic branch of the acoustic-gravity wave dispersion relationship. There is no resonance

TABLE II. Wave number k_{rc} and angle θ_c versus ϕ and I for an altitude of 600 km.

ϕ	0°	10°	20°	40°	60°	80°	90°
			$I = 7^o$, $\omega_c = 5.16\times10^{-4}$				
$k_{rc}^{(1)}$	6.884×10^{-6}	6.944×10^{-6}	7.127×10^{-6}	7.887×10^{-6}	9.269×10^{-6}	1.151×10^{-5}	1.310×10^{-5}
$\theta_c^{(1)}$	173.68	173.70	173.76	173.96	174.23	174.49	174.51
$k_{rc}^{(2)}$	4.641×10^{-5}	4.473×10^{-5}	4.039×10^{-5}	2.941×10^{-5}	2.070×10^{-5}	1.505×10^{-5}	1.310×10^{-5}
$\theta_c^{(2)}$	-174.96	-174.96	-174.95	-174.91	-174.83	-174.69	-174.51
			$I = 10^o$, $\omega_c = 1.05\times10^{-3}$				
$k_{rc}^{(1)}$	5.530×10^{-6}	5.598×10^{-6}	5.803×10^{-6}	6.664×10^{-6}	8.279×10^{-6}	1.109×10^{-5}	1.326×10^{-5}
$\theta_c^{(1)}$	165.88	165.96	166.19	166.98	167.90	168.70	169.01
$k_{rc}^{(2)}$	6.172×10^{-4}	2.250×10^{-3}	3.300×10^{-4}	6.162×10^{-5}	2.758×10^{-5}	1.629×10^{-5}	1.326×10^{-5}
$\theta_c^{(2)}$	10.210	10.210	-169.79	-169.75	-169.61	-169.27	-169.01
			$I = 20^o$, $\omega_c = 4.06\times10^{-3}$				
$k_{rc}^{(1)}$	4.461×10^{-6}	4.525×10^{-6}	4.723×10^{-6}	5.637×10^{-6}	7.674×10^{-6}	1.255×10^{-5}	2.110×10^{-5}
$\theta_c^{(1)}$	113.67	114.45	116.60	123.08	129.43	133.91	135.49
$k_{rc}^{(2)}$	9.443×10^{-6}	9.757×10^{-6}	1.078×10^{-5}	1.672×10^{-5}	5.861×10^{-5}	3.327×10^{-5}	2.110×10^{-5}
$\theta_c^{(2)}$	48.130	47.83	47.04	44.94	43.56	-136.18	-135.49
			$I = 40^o$, $\omega_c = 1.43\times10^{-2}$				
$k_{rc}^{(1)}$	1.257×10^{-5}	1.258×10^{-5}	1.259×10^{-5}	1.258×10^{-5}	1.218×10^{-5}	No Resonance	
$\theta_c^{(1)}$	81.72	82.59	85.28	97.45	125.86		
$k_{rc}^{(2)}$	1.150×10^{-5}	1.150×10^{-5}	1.150×10^{-5}	1.151×10^{-5}	1.156×10^{-5}	No Resonance	
$\theta_c^{(2)}$	173.29	173.18	172.82	170.94	163.86		

at some angle ϕ beyond 60°. Both $k_{rc}^{(1)}$ and $k_{rc}^{(2)}$ barely change with ϕ, and the directions of propagation are in the first and fourth quadrant for ϕ less than some value between 20° and 40°. For $\phi \geq 40^{\circ}$, waves are resonant in the fourth quadrant.

One may also inspect Table II for fixed ϕ as I increases. Most of the variations as described by using the diagrams in Figs. 2 to 4 are observable, such as the change of one of the resonant directions from third-quadrant angles to first-quadrant angles as I increases for a fixed ϕ. It should be noted that between $I = 20^{\circ}$ and $I = 40^{\circ}$ the resonant waves change from the gravity wave branch to the acoustic wave branch. For a range of values of I between 20° and 40° and for $\phi \leq 60^{\circ}$ the resonance must be cut off, since the dip angle lies between the tangent cases exhibited in Figures 3 and 4. However, this cutoff is exhibited only for $\phi = 80^{\circ}$ and $\phi = 90^{\circ}$ in the table, $I = 40^{\circ}$ lying in the cut-off range for these values.

D. Discussion

The sections above have demonstrated the nature of the two resonant directions that occur for a single frequency when a magnetic field is present and have shown how the resonances change with dip angle I and the angle ϕ, the angle of propagation from the meridian plane. The possibility of acoustic branch resonances is exhibited. It is most important to note that the calculated frequencies and directions for resonance turn out to be in the range of observed values for TID's as measured by incoherent scatter techniques, other ground measurements, and satellite measurements.[5,8-16] These results are considered significant and indicative of the possible connection of TID's and the resonance phenomenon.

It has been demonstrated that a strong resonance type of response may be possible in the F region at a particular frequency from a region that may be as great as 100 km in altitude. If this premise is confirmed by a full wave analysis, the resonance is far less localized in altitude than may have been expected. The resonance phenomenon also produces a strong response at each altitude over a relatively large frequency range, at least as much as 3:1 about the resonance frequency, and over a sizable range of directions that may be at least as large as 10° in width at lower altitudes and even much greater at higher altitudes.

Experimental verification of the coupled waves is essential. However, since an ionization acoustic wave excited by coupling has the same frequency and wave vector as the acoustic gravity wave perturbing the medium, it is difficult to distinguish the two waves by present methods that have been used to detect TID's. Nevertheless, it is of interest to analyze TID data on frequency and wave number statistically with respect to values expected for coupled waves. The simultaneous measurement of TID characteristics and atmospheric and ionospheric background parameters would aid considerably

in identifying the coupled wave. Unfortunately, such background data are not normally available, so that special experimental efforts are necessary. Experiments that take advantage of any of the characteristics of the ionization acoustic wave may also be conceivable, but further study of such possibilities is needed.

More theoretical work is necessary, particularly a full wave analysis to confirm whether the suggested strong nonlocalized response is real, with viscosity and thermal conductivity losses incorporated as well. When excitation is strong, nonlinear effects may enter, and such complications deserve treatment as well.

National Aeronautics and Space Administration
NGR 33-006-047

H. Eun and S.H. Gross

REFERENCES

1. H. Eun and S.H. Gross, "Ionospheric Disturbances and Gravity Waves," J. Geophys. Res., 81, 3261-3270 (1976).
2. C.O. Hines, "Electron Resonance in Ionospheric Waves," J. Atmos. Terr. Phys., 9, 56-70 (1956).
3. C.O. Hines, "Internal Atmospheric Gravity Waves at Ionospheric Heights," Can. J. Phys., 38, 1441-1480 (1960).
4. R.M. Clark, K.C. Yeh and C.H. Liu, "Interaction of Internal Gravity Waves with the Ionospheric F2-layer," J. Atmos. Terr. Phys., 33, 1567-1576 (1971).
5. J. Testud and P. Francois, "Importance of Diffusion Processes in the Interaction Between Neutral Waves and Ionization," J. Atmos. Terr. Phys., 33, 765-774 (1971).
6. C.O. Hines, "The Upper Atmoshpere in Motion," Geophys. Monogr. Ser., 18, American Geophysical Union, Washington, D.C. (1974).
7. S.H. Gross and H. Eun, "Traveling Neutral Disturbances," Geophys. Res. Lett., 3, 257-260 (1976).
8. T.M. Georges, "H F Doppler Studies of Traveling Ionospheric Disturbances," J. Atmos. Terr. Phys., 30, 735-746 (1968).
9. K. Davies and J.E. Jones, "Three-dimensional Observations of Traveling Ionospheric Disturbances," J.Atmos. Terr. Phys., 33, 39-46 (1971).
10. P.L. Dyson, G.P. Newton and L.H. Brace, "In Situ Measurements of Neutral and Electron Density Wave Structure from the Explorer 32 Satellite," J. Geophys. Res., 75, 3200-3210 (1970).
11. J.P. Friedman, "Propagation of Internal Waves in a Thermally Stratified Atmosphere," J. Geophys. Res., 71, 1033-1054 (1966).
12. C.S.G.K. Setty, A.B. Gupta and O.P. Nagpal, "Ionospheric Response to Internal Gravity Waves Observed at Delhi," J. Atmos. Terr. Phys., 35, 1351-1361 (1973).
13. J. Testud, "Gravity Waves Generated During Magnetic Substorms," J. Atmos. Terr. Phys., 32, 1793-1805 (1970).
14. G. Thome, "Long-period Waves Generated in the Polar Ionosphere During the Onset of Magnetic Storms," J. Geophys. Res., 73, 6319-6336 (1968).
15. G. Vasseur and P. Waldteufel, "Thomson Scatter Observations of a Gravity Wave in the Ionospheric F-region," J. Atmos. Terr. Phys., 31, 885-888 (1969).
16. K.C. Yeh and C.H. Liu, "Acoustic Gravity Waves in the Upper Atmosphere," Rev. Geophys. Space Phys., 12, 193-216 (1974).

A KELVIN-HELMHOLTZ TYPE INSTABILITY IN PLASMA STRIATIONS

H. Kudyan and K. Chung

It was reported previously that plasma striations are closely associated with low-frequency weak plasma turbulence and strong ion beams along the confining magnetic field.[1] Recognizing the role of ion beams in splitting plasma sheets and inducing distinct spectral peaks, we have further carried out experimental investigations on the weak plasma turbulence caused by an ion-beam induced instability.

Although the cross-sectional profile of a magnetically confined electron cyclotron resonance plasma may take a large number of different apparent structures, there exist fundamentally two types of profiles: hollow (cylindrical) column and fully-developed-column. Hollow column plasma is very much susceptible to breaking into striations and usually exhibits plasma instability which develops into weak plasma turbulence as striations form. The striated plasma column may actually be termed as the fully developed weakly turbulent state of hollow column plasma. On the verge of breaking up, hollow column plasma exhibits very strong oscillations caused by a plasma instability.

Using a phase-sensitive lock-in amplifier loop for measuring the wave-pattern in the plasma and employing concentric electrodes to control the axial ion-beams, we have observed the frequency variation of the instability and the growth-decay behavior of the instability in helium, nitrogen, air and argon discharges. In addition, we have measured the distribution functions using small electrostatic analyzers[1] and the plasma potential by electrostatic probes.

Until the instability develops into weak turbulence, we observe standing wave behavior in both the radial and axial directions. Such standing wave patterns deteriorate as the plasma becomes weakly turbulent and breaks into striations. Whereas the oscillation frequency is roughly in proportion to the background potential radial gradient, it is virtually independent of the mass of ions and the magnitude of the confining magnetic field. In addition, it is not possible to match the observed frequency with either the ion acoustic wave frequency or ion cyclotron frequency based on ranges of parameters of our plasma. The frequency range, 20 ~ 90KHz, indicates strongly low-frequency electrostatic instability as the cause of the oscillations.

In view of the strong ion beams in the unstable plasma, we have made a detailed measurement of the flow pattern of ion beams. In Fig. 1, we show a typical measurement of the radial potential distribution, the relative flow distribution and the energy analyzer data in a striated plasma column. (The picture is taken axially to show the plasma striations.) As suspected, the plasma is imbedded with flow shears both in azimuthal and axial directions. When we apply a potential (via concentric grids or

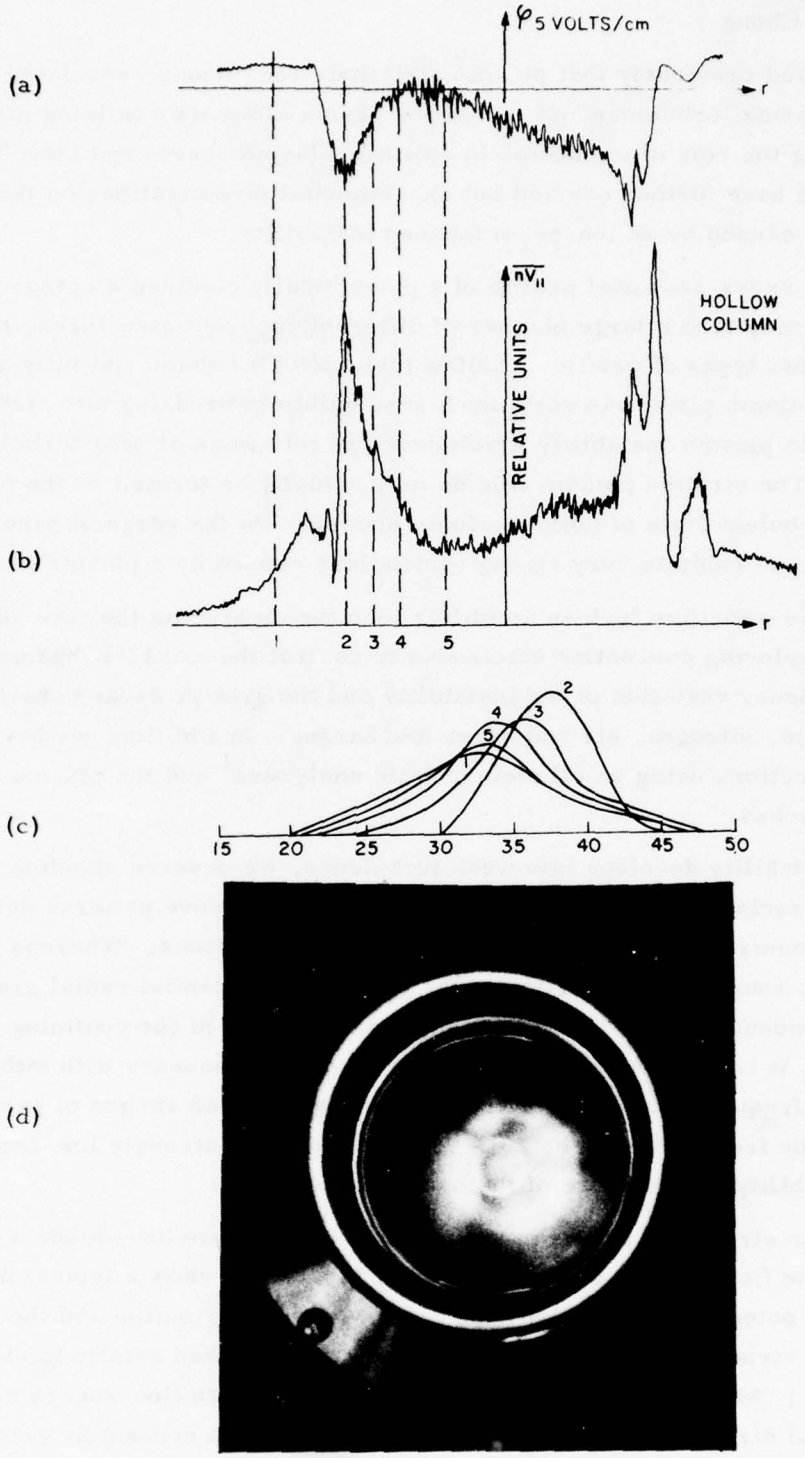


Fig. 1. (a) Radial profiles of plasma potential, (b) axial drift and (c) distribution function, (d) photograph taken axially to show plasma striations.

electrodes), we are able to ignite the splitting of plasma column or subdue the plasma oscillation by reversing the polarity of the applied potential. It is to be noted that the application of a radial potential modifies both the axial and radial flow patterns. This is due to the fact that the sum of conservative (eE, ∇p) and non-conservative forces must balance each other in an equilibrium plasma. This situation induces shear fields in the plasma column, which have pronounced effects on a Kelvin-Helmholtz type instability. The oscilloscope display shows the quenching and growing of the Kelvin-Helmholtz type instability observed in the striated plasma (see Figure 2).

The electrostatic Kelvin-Helmholtz type instability was previously observed in Q-machines,[2,3] in which shear flow fields were created either by imposed mechanical walls or by plasma rotations. Our observation implies, however, that the similar Kelvin-Helmholtz type instability is capable of splitting plasma sheets into striations as the instability grows into weak plasma turbulence. This is a rather serious implication since plasma striations were observed in plasma clouds created in a flowing neutral background and their charge sheets were subject to splitting when conditions deviate from the uniformity. One must be aware of the possibility of the formation of striations when the plasma is subject to strong flow fields and shears of the flow fields develop.

Joint Services Technical Advisory Committee
F44620-74-C-0056

H. Kudyan and K. Chung

REFERENCES

1. H. Kudyan and K. Chung, "Velocity Distribution and Fluctuation Spectra in a Weakly Turbulent Plasma," Progress Report No. 40 to JSTAC, Polytech. Inst. of New York, Report No. R-452.40-75, 276-281 (1975).
2. N. D'Angelo and S. VonGoeler, "Investigation of the Kelvin-Helmholtz Instability in a Cesium Plasma," Phys. Fluids, 9, 309 (1966).
3. G.I. Kent, N.C. Jen and F. Chen, "Transverse Kelvin-Helmholtz Instability in a Rotating Plasma," Phys. Fluids, 12, 1240 (1969).

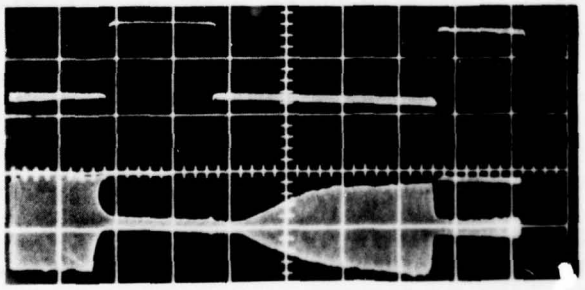


Fig. 2. Oscilloscope display showing the decay and growth of the instability.

THERMAL ANALYSIS OF VACUUM VESSEL BAKE-OUT OF FUSION TEST REACTOR

K. Chung

The torus vacuum vessel of a Tokamak-type fusion reactor should be prepared for the very high level vacuum operation. Although the individual components are pre-baked prior to assembly, it is normally required to perform in-situ bakeout of the vacuum vessel in order to remove impurities from the inner surfaces.

There exist several conditions and constraints for the in-situ bake-out process:

(1) The vacuum vessel must be heated to a temperature above 250°C (482°F). Depending on the vacuum performance of the vessel, the required bake-out temperature can be as high as 500°C.

(2) Due to the mechanical stress consideration, the temperature differences in the vacuum vessel should be maintained less than 50°C during both the steady-state and transient state.

(3) In view of the normal operation requirements, the heating up time should be limited within approximately four hours and the steady-state bake-out time may be as long as several days.

(4) The bake-out process must be repeatable under remote maintenance conditions, since the vacuum vessel will become radioactive after D-T operation.

(5) The hardware required for the in-situ bake-out should be comprised with the other structures of the test reactor and the bake-out system should not impair the performance of the device.

There exists several alternatives for the bake-out heating:

(1) Hot gas heating: the vacuum vessel is heated by flowing hot gas (helium, argon or air) through heat pipes surrounding the vacuum vessel. The inner surfaces of the vessel are heated through heat transfer from the heat pipes.

(2) Hot liquid heating: to increase the heat capacity (and thus reduce the flow rate and temperature gradient along the heat pipe) hot organic fluid can be used as the medium. Again the heat transfer rates determine the temperature profiles of the vacuum vessel.

(3) Heat blanket: as in normal bake-out processes the vessel is surrounded by heat blankets. Heating is done from the outer shell and the inner shell and structures are heated via conduction and radiation.

(4) Electrical resistive heating: passing electrical currents (ac or dc), the vacuum vessel is heated to the bake-out temperature. The electrical current is supplied by the external power sources via terminals attached to the vacuum vessel.

(5) Inductive heating: change of the toroidal magnetic field induces poloidal currents in the vacuum vessel. Ohmic heating due to the induced poloidal currents can be used for heating the vacuum vessel for bake-out.

(6) Plasma discharge heating: large amount of energy can be deposited to the inner walls of the vacuum by the radiation and particle loading from the discharges. Discharge cleaning is itself an effective method for vacuum preparation of the inner walls of the vessel. It can be also used to heat the vacuum vessel for the bake-out process.

Assessing these alternatives against the desired conditions and constraints and further comparing them with respect to spatial requirements, structural simplicity and economics, hot gas heating and electrical resistive heating are chosen as the primary candidates.

Inductive heating is considered as the supplementary system. Hot gas heating is attractive since it can be combined with the vessel cooling system. For heating bellows and internal structures directly and in a programmed way, the electrical resistance heating provides the most flexible scheme. Since the induction heating can be applied without any additional equipment, it is available for application. Thus, for the in-situ bake-out process of the vacuum vessel, hot gas heating and/or electrical resistive heating are the acceptable solutions with induction heating as a backup operation.

The design of the sizes and layout of the heat sources depends on the thermal characteristics of the vacuum vessel. Because of complicated shapes and boundary conditions, the thermal characteristics of the vacuum vessel can be determined only by the numerical solution of the time-dependent heat conduction (radiation convection) equation which includes heat sources externally conducted or internally generated. For this computation, there exist several computer codes and we have chosen LION program developed by Knolls Atomic Power Laboratory for CDC computers. In addition to the solving of heat conduction in structure elements, LION may also be used to solve problems involving forced convection, free convection, or radiation for finding temperatures and heat fluxes on the surface. The LION program utilizes a first forward difference method based on a nodal representation of the geometry through the evaluation of the equivalent resistances of the nodal connectors.

The TFTR vacuum vessel is made of 305 Stainless Steel whose thermal conductivity, specific heat and electrical resistivity are 10.2 Btu/hr/ft^2/ft/$^\circ$F (at 392°F), 0.12 Btu/lb/$^\circ$F and 72 $\mu\Omega$-cm. The physical shape and thermal analysis model are described in Figure 1. The vacuum vessel is composed of sectionized solid wall bellows combinations which are connected along the major axis of the torus. Bellows, whose function is to increase the toroidal resistance of the vacuum vessel, are also made of thin stainless steel sheets and protected from both inside against the hot plasma and outside against pressure. The thickness and detailed structure of the vacuum vessel are determined from other design requirements such as stress, conductivity, maintenance, etc. Once the configuration is given, we may determine the required level of the heat generation for the bake-out heating using the LION code. An example of the computation is given in Figure 2, which shows the time dependent temperature profile of the vacuum vessel for electrical resistance heating of 5.0 Btu/hr-in^3 in solid walls and 0.5 Btu/hr-in^3 in bellows. The result is acceptable. This particular heat generation rate can be achieved

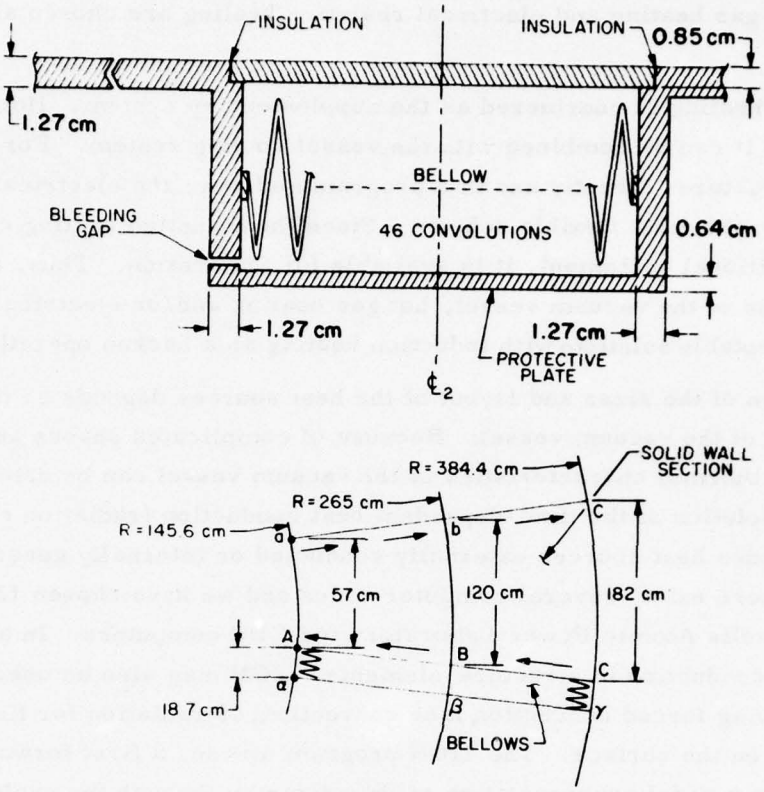


Fig. 1(a). Physical shape of TCT vessel.

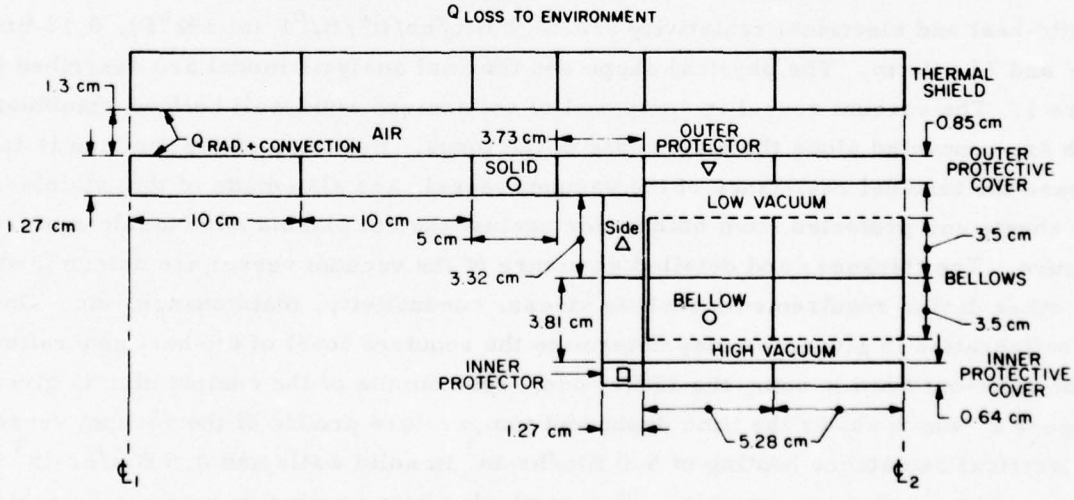


Fig. 1(b). Thermal analysis model.

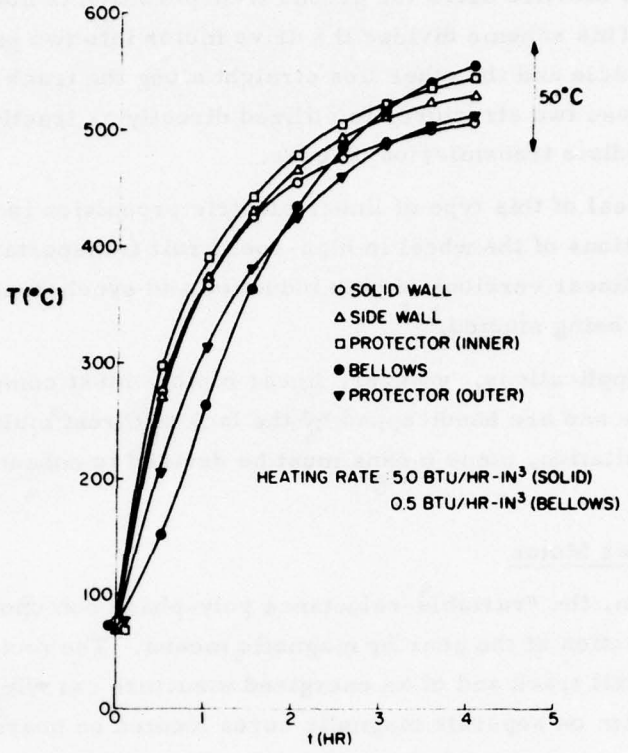


Fig. 2. Bake-out resistive heating.

by toroidal electric current of ~.8 kA and poloidal electric current (in the solid wall) of ~10 kA. Such electric currents can be easily applied by external power sources.

The final design of the bake-out heating system depends on more than the thermal analysis of the vacuum vessel alone. It is required to incorporate other design considerations such as cooling system during the discharge cleansing, needs for structural reinforcement, protection against disruptive discharges, etc.

Ebasco Services Inc.
NY 709002 under Prime E(11-1)-3073

K. Chung

NEW CONCEPTS IN LINEAR PROPULSION

E. Levi

A novel type of electric drive for ground transportation is now undergoing intensive development. This scheme divides the drive motor into two parts, of which one is carried by the vehicle and the other lies straight along the track. The force of interaction between these two structures is utilized directly as tractive effort, without the need for intermediate transmission or gear.

The major appeal of this type of linear electric propulsion is that it promises to overcome the limitations of the wheel in high-speed rail transportation. For high-speed applications, linear versions of both induction and synchronous motors of conventional design are being studied.[1-5]

In low-speed applications, however, linear motors must compete with their rotating counterparts and are handicapped by the lack of thrust multiplying gears. Hence, to overcome this limitation, some means must be devised to enhance the thrust developed per unit weight.

A. An Unusual Linear Motor

A new invention, the "variable-reluctance poly-phase homopolar converter," accomplishes the function of the gear by magnetic means. The motor consists of a passive ferromagnetic rail track and of an energized structure carrying the phase windings of a poly-phase system on separate magnetic cores located on board the vehicle.

Figure 1 is a photograph of a small-scale, but fully operational, model of the motor, while Fig. 2 is a perspective view of one phase winding, its magnetic core, and the corresponding section of the passive rail track.

The thrust develops as a result of the tendency of the teeth in the energized structure to align themselves with those of the rail track. The phase cores are solidly mounted together and spaced from one another a distance corresponding to the phase angle of the current they carry, so that their combined effect is a traveling wave.

The main features of the motor are:

(1) The thrust is proportional to the number of teeth and can be made practically independent of the weight of the motor

(2) A single winding carries both field and armature excitation resulting in high efficiency and low weight of the part of the motor carried by the vehicle

(3) The motor develops significant forces of attraction between the energized structure and the passive rail track -- forces which can be utilized for magnetic levitation

(4) The low-profile rail track shown in Fig. 1 makes this propulsion system compatible with existing rail systems

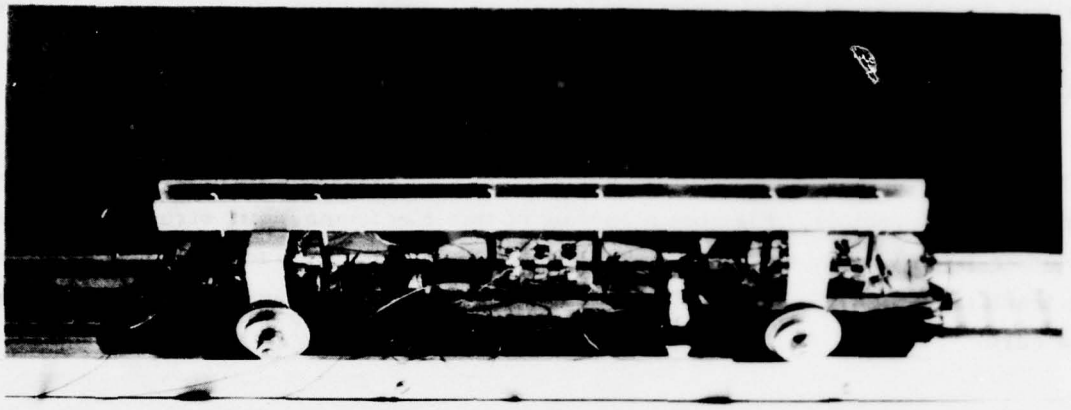
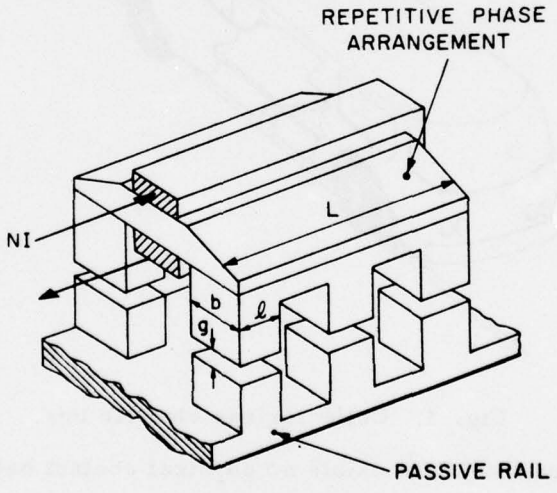


Fig. 1. Model of novel propulsion system for light-rail transit.



NOTE: Proposed design does not use raised teeth on rail (i.e., in-ground strip).

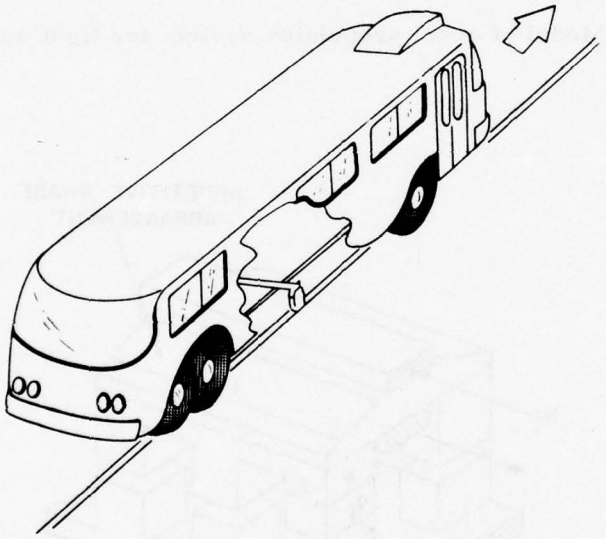
Fig. 2. Variable reluctance polyphase homopolar converter.

Since high values of thrust weight ratios are easily attainable, this motor is particularly suited to relatively low-speed/high tractive effort applications, such as rapid transit and light-rail transit.

B. A Collectorless Electric Bus

Another invention, the collectorless electric bus, is aimed at overcoming the serious drawbacks of buses propelled by internal combustion engines, i.e., (a) low efficiency, (b) consumption of critical energy resources, (c) noise, and (d) chemical pollution.

The propulsion system consists of a low-voltage energized track which is emplaced in the road pavement, flush with its surface, and a passive electromagnetic structure carried by the vehicle. Flexible mounting of this electromagnetic structure on board a rubber-tired vehicle would allow its lateral displacement up to a full bus width from the track (see Figures 3 and 4). As a result the bus can maneuver in traffic and pull to a curb.



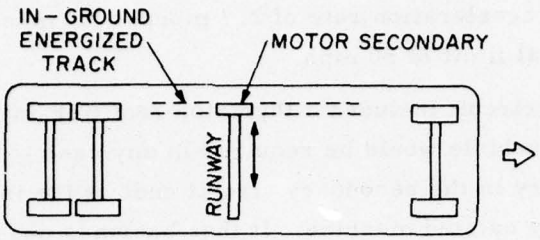
Linear electric motor

NOTE: Cut-away shows motor secondary mounted on a below vehicle runway for lateral mobility.

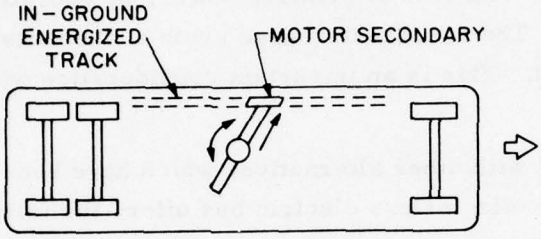
Fig. 3. Collectorless electric bus.

It should be noted that there exists no physical contact between the power source and the vehicle. The power transfer is via induction. The gap separating the vehicle and roadbed parts of the motor is approximately half an inch (see Figure 4). The forces of attraction between the primary and secondary structure are utilized to maintain the necessary alignment and guide the movable part along the fixed one.

The structurally simple and sturdy primary consists of straight conductors laid parallel along the direction of motion, and intertwined with laminated iron cores. The view from below in a small-scale model is shown in Figure 5.



(a) Design 1: electric motor secondary sliding on runway.



NOTE: Permitted degrees of motion shown by arrows.

(b) Design 2: electric motor secondary with swinging arm, arm sliding through rotating fixed head.

Fig. 4. Two designs for mounting of propulsion system on collectorless electric bus.

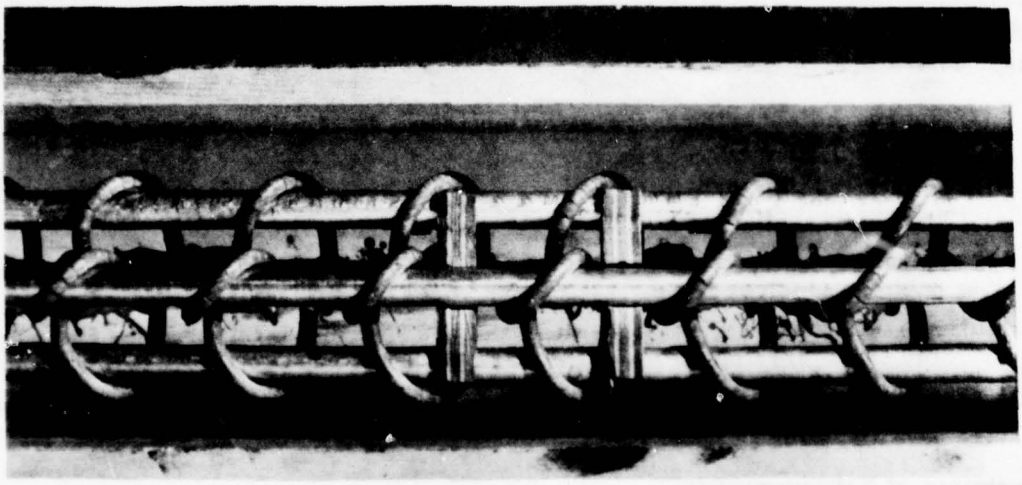


Fig. 5. Aluminum conductors and ferro-magnetic cores viewed from underside.

Preliminary calculations have been made for a typical vehicle with a passenger capacity of 53 seats and 40 standees. Two-inch diameter, aluminum conductors will suffice to provide an acceleration rate of 2.7 mphps (1.2 m/sec^2) and speeds well in excess of the practical limit of 50 mph.

The secondary circuit includes a thyristor rectifier-inverter and a battery. A battery on board the vehicle would be required in any case to provide auxiliary power. Insertion of the battery in the secondary circuit endows the induction motor with the properties of a doubly excited machine. It thus becomes possible to control the speed, the acceleration and deceleration, and even the direction of energy flow, exclusively from the vehicle side. Moreover, forward and backward motion, as well as braking can be achieved, even with loss of primary power, by operating the motor as a variable reluctance machine. The charge-discharge cycle of the battery accomplishes a leveling of the peak in demand. This is an important consideration with regard to energy conservation.

When compared with other alternatives which have been considered for the existing urban buses, the collectorless electric bus offers the following advantages:

(1) It can maneuver up to about eight feet laterally, a feat which tracked vehicles, such as trolley cars cannot perform

(2) It needs no unsightly overhead wires from which trolley-buses inconveniently become un-tracked

(3) It requires no technological break-throughs, as is the case with electric battery and kinetic energy buses.

Moreover, the propulsion system for the collectorless electric bus, as well as the variable-reluctance poly-phase homopolar converter described above may permit the realization of many of the advantages of linear motors over their rotating counterparts. These are:

(1) Increased adhesion force with consequent feasibility of faster acceleration and deceleration

(2) Improved ride dynamics and comfort (of benefit to elderly/handicapped riders)

(3) Regenerative braking and, in general, higher efficiency

(4) Greater safety resulting from the presence of an additional electric braking system which does not rely on the wheel or rail

(5) Noise abatement, as a result of the absence of gears, reduced flat wheels (no tread braking or locked wheel sliding) and reduction in rail wear and tear

(6) Lower maintenance costs of propulsion and braking systems and tracks

(7) Lower pollution from brake shoes (cast iron, asbestos).

Department of Transportation
DOT-FR-64227

E. Levi

REFERENCES

1. E. Levi, "Linear Propulsion," Sec. 23, Par. 269-307, Eleventh Edition, Standard Handbook for Electrical Engineers (New York: McGraw Hill, 1977).

2. R.B. Powell, "Linear Induction Motor Electrical Performance Test," FRA-OR ED 76-265 (June 1976).

3. E. Levi, "A Preliminary Evaluation of Electrical Propulsion by Means of Iron-Cored Synchronously Operating Motors," Polytechnic EE/EP Report No. 75-011 (January 1975).

4. E. Levi, "Preliminary Design Studies on Iron-Cored Synchronously Operating Linear Motors," Polytechnic EE/EP Report No. 76-005 (February 1976).

5. C.A. Skalski, "The Air-Cored Linear Synchronous Motor - An Assessment of Current Developments," MITRE Technical Report, MTR-7028 (September 1975).

MAGNETOSTATIC POTENTIAL AND MAGNETIC FLUX DENSITY ALONG THE ARMATURE OF HOMOPOLAR INDUCTOR MACHINES

E. Levi and M.S. Gemelos

A new method for the determination of static field distributions in complex geometries and with arbitrary field distributions was developed in previously reported studies.[1,2] This method was first applied to homopolar inductor type machines, under the assumption that the armature surface is smooth and that the armature excitation consists of a sinusoidal current sheet. In reality the armature conductors are placed in discrete slots and the examination of iron saturation effects requires a more detailed description of the physical structure.

The aim of this study is the determination of the magnetic flux density and of the magnetostatic potential along a plane lying on the armature tooth tips. The problem is assumed to be two-dimensional, the current carrying conductors are removed and the slots assumed to have infinite depth, so that conformal mapping procedures and in particular the Schwarz-Christoffel transformation can be used. The armature teeth are assigned appropriate potentials with respect to the pole face located on the other side of the air gap.

To gain more insight, a step by step approach was adopted. First a simple configuration is examined containing an isolated slot. Then, more involved schemes, such as a succession of open slots are studied. Taking advantage of the fact that linearity holds in the air gap, the field and armature excitation are also examined separately.

The simplest case is the classical Carter problem[3] of an isolated slot with field excitation only. The complicating factor is that interest is focused here on the flux distribution at the surface of the slotted armature, rather than at the smooth surface across the gap. It is found that the perturbation caused by the presence of the slot dies out a short distance away along the gap.

The case of adjacent slots, also with field excitation, yields an exact solution in terms of the Jacobian elliptic functions. The flux density and the potential along the armature surface are given in Figure 1. A remarkable result is that in the case studied in which the ratio of gap length over the slot width is two, the field under the slot is as strong as the field under the tooth face. This is due to the large ratio of gap length over slot width. When the slot is very deep, changes in the gap length cannot affect significantly the field strength under the slot, while the field under the tooth is strongly affected by these changes.

With armature excitation and in contrast with the case of field excitation, it is found that the perturbation caused by an isolated slot dies away slowly along the gap. The interaction between adjacent slots then, becomes important.

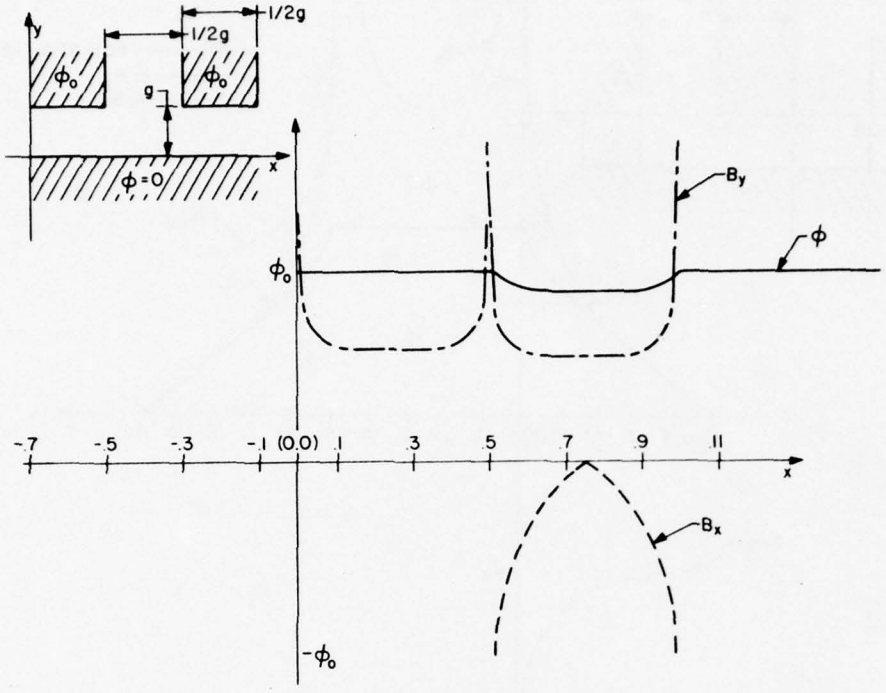


Fig. 1. Field excitation

With a succession of teeth at different potentials a rigorous solution becomes too cumbersome because of the lack of symmetry. However, since the field is required only along the armature surface and not within the gap and since the armature excitation results in a discrete stepwise increase of the magnetostatic potential from one tooth to the adjacent one, a solution can be obtained by treating only a couple of slots and by using superposition.

Plots of the potential and B-field along the armature surface for the elementary two-slot configuration are shown in Figure 2. It is found that the perturbation caused by the slot dies out faster than in the case of an isolated slot. This results in a more uniform field under the tooth face, albeit at a lower value than the asymptotic one.

Finally, the typical case of a succession of three slots belonging to the same phase winding is shown in Figure 3. The potential and B-field distributions are obtained by superimposing the plots of Fig. 2, multiplied by the factors 2, 1 and -1 over the regions of teeth at potentials $2\,\phi_0$, ϕ_0 and $-\phi_0$, respectively. The pole surface is taken at zero potential.

These results are in good agreement with plots obtained by analogue simulation using a Pasco Scientific Equipotential and Field Gradient Mapper.

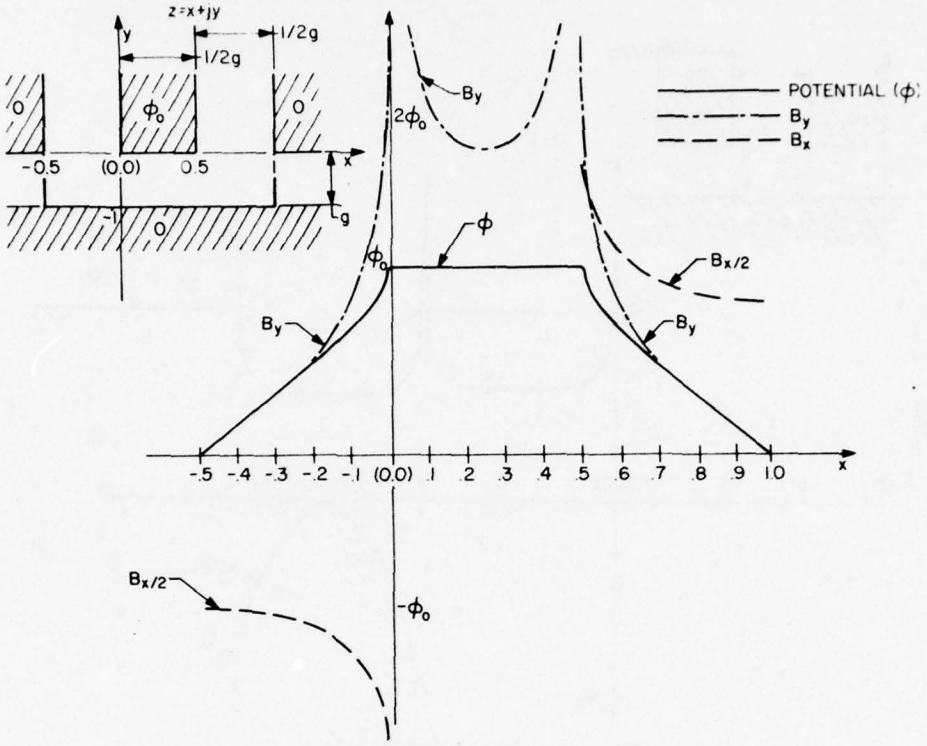


Fig. 2. Armature excitation (for two slots).

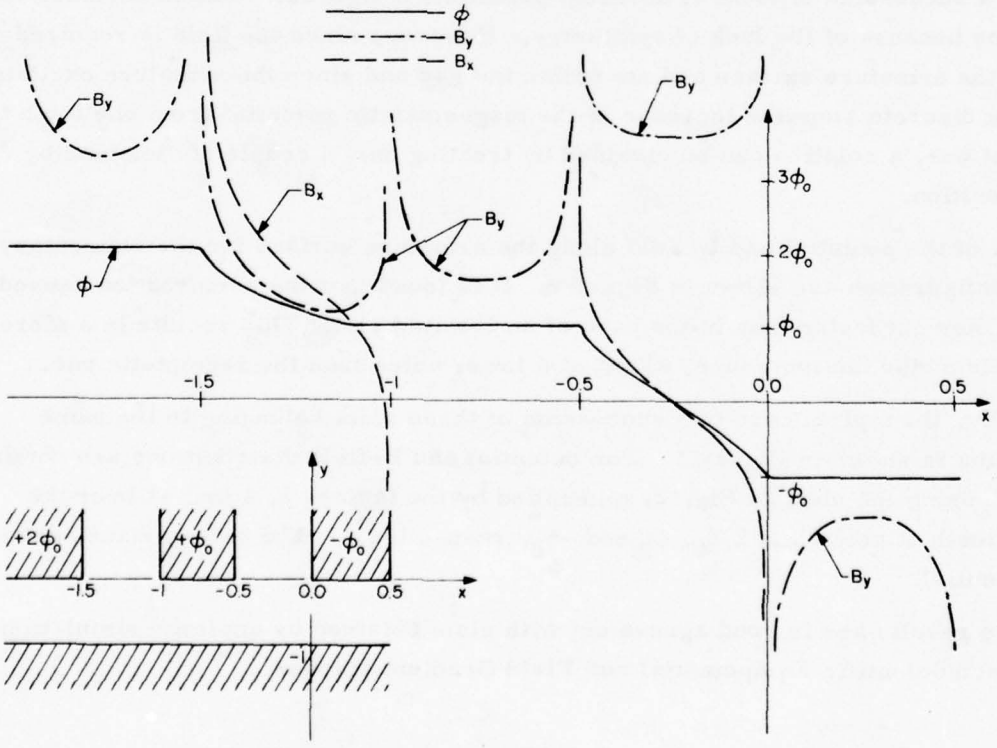


Fig. 3. Armature excitation (for a succession of slots).

In further studies iron saturation effects will be taken into account by modifying the magnetostatic potentials in an iterative procedure.

Department of Transportation
DOT-FR-64227

E. Levi and M.S. Gemelos

REFERENCES

1. E. Levi and J.P. Lee, "A New Method for the Determination of Magnetic Field Distributions," Progress Report No. 40 to JSTAC, Polytech. Inst. of New York, Report No. R-452.40-75, 307-313 (1975).

2. E. Levi, "Preliminary Design Studies on Iron-Cored Synchronously Operating Linear Motors," Research Report No. POLY-EE/EP 76-005, Polytech. Inst. of New York (May 1976).

3. F.W. Carter, Journal IEE, 29, part 146, 925 (1900).

NEW RESULTS IN ELECTRICAL MACHINE THEORY: FORMULAS OF THE BLASIUS TYPE FOR THE CALCULATION OF FORCE AND TORQUE ON THE ROTOR OF A CYLINDRICALLY UNIFORM ELECTROMAGNETIC MACHINE VIEWED AS AN AEROFOIL IMMERSED IN A MAGNETIC FLUID

D.C. Youla

Consider a magnetized current-carrying ferromagnetic body separated from all other sources by a free-space region. As is well known,[1,2] the expressions

$$\underset{\sim}{F} = \frac{\mu_o}{2} \int_{S'} H^2 \underset{\sim}{t}\, da \tag{1}$$

and

$$\underset{\sim}{T} = \frac{\mu_o}{2} \int_{S'} H^2 \underset{\sim}{r} \times \underset{\sim}{t}\, da \tag{2}$$

give the correct values for the total external magnetostatic force and torque exerted on the complete body provided the orientation of the unit vector $\underset{\sim}{t}$ with respect to the magnetic intensity $\underset{\sim}{H}$ and the positive normal $\underset{\sim}{n}$ erected at da is as shown in Figure 1. (S' is any closed surface in the source-free region which encloses the body, $H = |\underset{\sim}{H}|$ and $\underset{\sim}{r}$ is the radius vector of the element of area da from the origin of coordinates. The surface S' can be chosen coincident with S if it is understood that $\underset{\sim}{H}$ is the magnetic intensity just above da and outside of V.)

Let us now assume that the problem is 2-dimensional and the body possesses uniform cross-section in the z-direction. Then $\underset{\sim}{H} = \underset{\sim}{H}(x,y)$ and $\underset{\sim}{k} \cdot \underset{\sim}{H} = 0$. In the source-free region $\nabla \times \underset{\sim}{H} = \underset{\sim}{O}$ and there exists a scalar function u(x,y) such that

$$\underset{\sim}{H} = -\nabla u(x,y) \quad . \tag{3}$$

Since $\nabla \cdot \underset{\sim}{H} = 0$,

$$\frac{\partial H_x}{\partial x} + \frac{\partial H_y}{\partial y} = 0 \tag{4}$$

and

$$\nabla^2 u(x,y) = 0 \quad . \tag{5}$$

Thus u(x,y) is harmonic (but not necessarily single-valued). Along a streamline of $\underset{\sim}{H}$,

$$H_y dx - H_x dy = 0 \quad . \tag{6}$$

However, as a consequence of Eq. (4),

$$H_y dx - H_x dy = \text{exact differential} = dv(x,y) = 0, \tag{7}$$

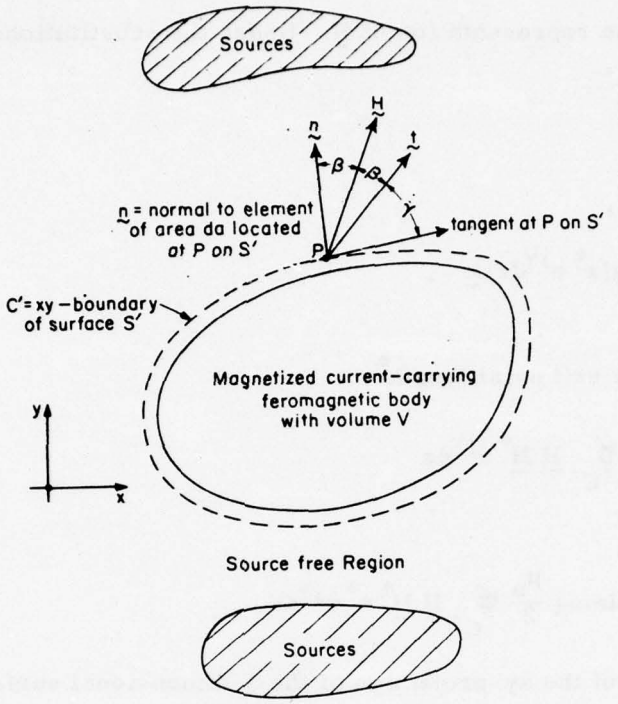


Fig. 1

v(x, y) a scalar function. Hence, the conjugate function v(x, y) is constant along a streamline of $\underset{\sim}{H}$ and

$$\frac{\partial v}{\partial x} = H_y = -\frac{\partial u}{\partial y} \quad , \qquad (8)$$

$$-\frac{\partial v}{\partial y} = H_x = -\frac{\partial u}{\partial x} \quad .$$

But these are precisely the Cauchy-Riemann equations which connect the real and imaginary parts of the analytic function $W = u + jv$ of the complex variable* $z = x + jy$. Since W(z) is analytic,

$$\frac{dW}{dz} = \frac{\partial u}{\partial x} + j\frac{\partial v}{\partial x} = \frac{\partial u}{\partial x} - j\frac{\partial u}{\partial y} \qquad (9)$$

$$= -(H_x - jH_y) = -\underline{H}^* \qquad (10)$$

---

*This z should not be confused with the space variable. As usual, a^* denotes the complex-conjugate of a.

where

$$\underset{\sim}{H} = H_x + jH_y \tag{11}$$

is the complex-plane representation of $\underset{\sim}{H}$. Under the substitutions

$$\begin{aligned} &\underset{\sim}{r} \rightarrow z \quad , \\ &d\underset{\sim}{r} \rightarrow dz \quad , \\ &\underset{\sim}{t}\,da \rightarrow e^{j\gamma}dz \quad , \\ &\underset{\sim}{r} \times \underset{\sim}{t}\,da \rightarrow \mathrm{Imag}(z^* e^{j\gamma}dz)\,\underset{\sim}{k} \quad , \\ &\underset{\sim}{n}\,da \rightarrow jdz \quad , \end{aligned} \tag{12}$$

it is evident that for unit axial length*

$$\underset{\sim}{F} \rightarrow \underset{}{F} = -\frac{\mu_o}{2} \oint_{C'} \underline{H}\,\underline{H}^* e^{j\gamma}dz \tag{13}$$

and

$$\underset{\sim}{k} \cdot T \rightarrow T_\varphi = -\mathrm{Imag}\,\frac{\mu_o}{2} \oint_{C'} \underline{H}\,\underline{H}^* z^* e^{j\gamma}dz \quad . \tag{14}$$

(C' is the boundary of the xy-projection of the 3-dimensional surface S'.) From the geometry of Fig. 1,

$$e^{j\gamma}\underline{H}^* dz = -j\underline{H}(dz)^* \tag{15}$$

whence,

$$\underline{F}^* = -\frac{j\mu_o}{2} \oint_{C'} \left(\frac{dW}{dz}\right)^2 dz \tag{16}$$

and

$$T_\varphi = \mathrm{Re}\,\frac{\mu_o}{2} \oint_{C'} \left(\frac{dW}{dz}\right)^2 z\,dz \quad . \tag{17}$$

Consequently, all force and torque calculations in cylindrically uniform ferromagnetic systems can be carried out with the aid of the same Blasius formulas used so effectively in the theory of aerofoils.[3] Furthermore, and this is very significant from a practical point of view, Eqs. (16) and (17) are valid even for saturated and non-infinitely-permeable

---

*Re a and Imag a stand for the real and imaginary parts of a, respectively.

iron. It should now be possible to attack the problem of designing optimal rotor profiles with the aid of the theory of residues. Any new results in this direction will be reported at the appropriate time.

Joint Services Technical Advisory Committee
F44620-74-C-0056

National Science Foundation
ENG 74-13054

D.C. Youla

REFERENCES

1. J.A. Stratton, "Electromagnetic Theory," (New York: McGraw Hill, Inc., 1941).
2. W.F. Brown, Jr., "Magnetoelastic Interactions," Springer Tracts in Natural Philosophy, 9 (1966).
3. H.Glauert, "The Elements of Aerofoil and Airscrew Theory," Second Edition, Cambridge University Press (1948).

STATIC TESTS ON THE CLAW POLE LINEAR SYNCHRONOUS MOTOR

E. Levi, L. Birenbaum, P. Chan, D. Fontana, F. Lalezari, J. Mullen, N. Rokkos, D. Wong

Static tests on the Nadyne motor are possible because the motor is of the synchronous type i.e., static tests on the linear induction motor are not possible because the developed forces are due to the induced currents. An idealized sketch of the motor and a typical magnetic flux-path are shown in Figure 1.

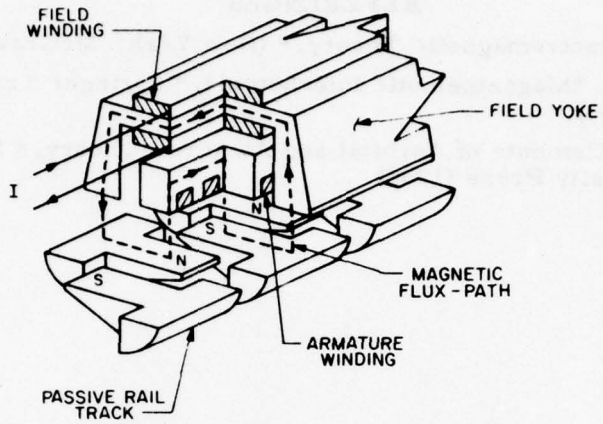


Fig. 1. Idealized sketch of claw-pole linear synchronous motor.

The field winding (on the vehicle section) is excited with direct current, which produces alternate north and south magnetic field poles in the part of the track underneath the vehicle. In order to perform static tests on the motor, the armature coils also are excited with direct current. During normal operation, polyphase AC would be used. Because of the way the armature coils are wound, a suitable choice of DC currents to feed the separate phases gives rise to a static magnetic field varying sinusoidally (approximately) along the armature. Figure 2 is a schematic diagram showing the vehicle core used for the model, including important dimensions. Figure 3 is a photograph of the salient pole track structure. The "claw" type interdigital arrangement of the poles is clearly visible. The interaction of the fields created by currents in the field winding and in the armature winding results in a horizontal thrust and a vertical levitation force.

The main purpose of the tests now to be described was to measure the magnetic fields in the air gap of the motor, and to measure the corresponding thrust and lift forces caused by their interaction.

A. Details of the Force Measurements

The magnetic fields were measured using a Bell gaussmeter. In order to measure

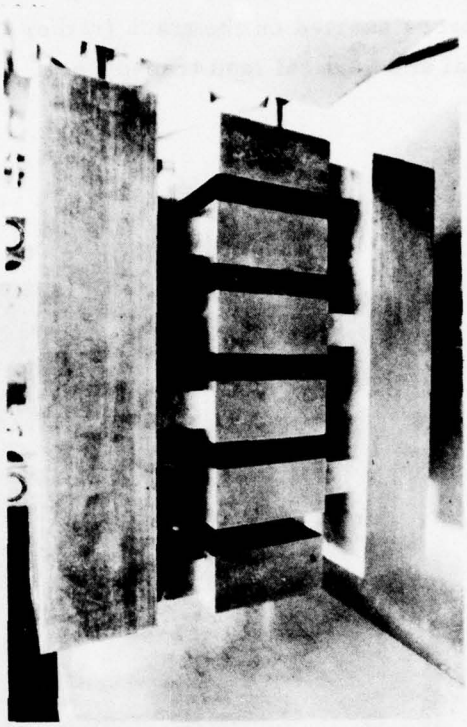


Fig. 3. View of the salient pole track structure.

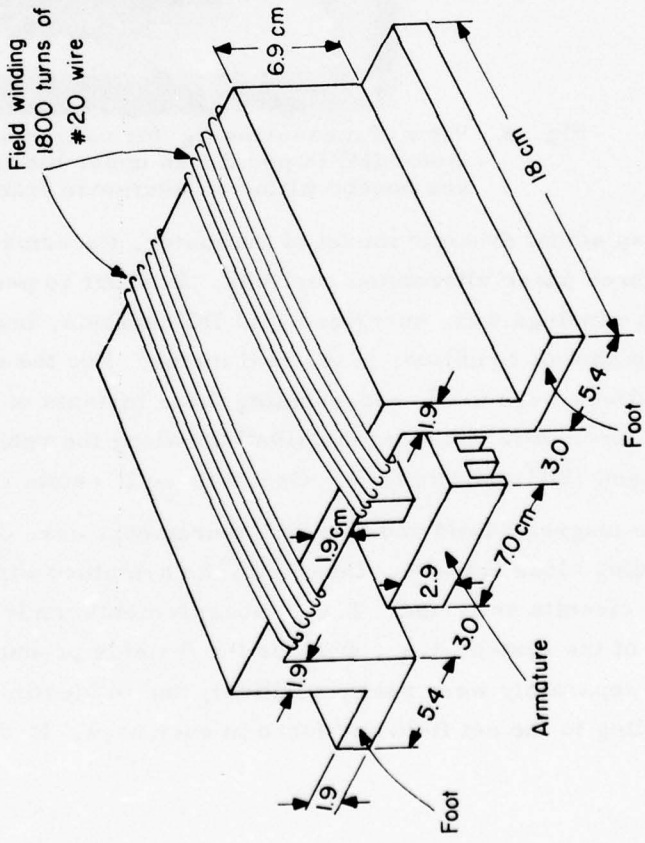


Fig. 2. Sketch of nadyne static model, showing vehicle core.

the horizontal and vertical forces, the motor was suspended in an aluminum retaining frame (Fig. 4) and the forces exerted on the track (rather than on the motor) were measured using horizontal and vertical load transducers.

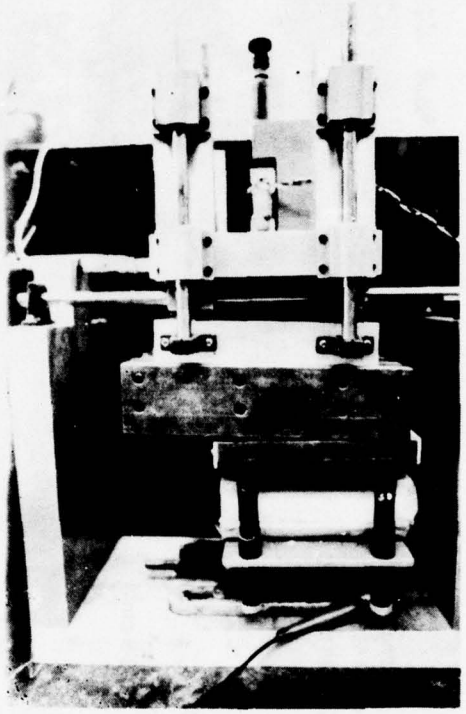


Fig. 4. View of measuring jig for nadyne static model. The motor (M) is positioned under the track (T). Both are housed within an aluminum frame.

In an actual dynamic model of the motor, the armature winding would be energized with three phase alternating currents. In order to perform the static tests, the armature windings were energized with DC currents, instead, to simulate an instantaneous snap shot of conditions in the real motor. For the measurement, three different configurations were used, representing three instants of operation, and intended to simulate three different m mf distributions along the vehicle. These three conditions were chosen 120° apart in time. One of these is shown in Figure 5.

The magnetic field and force measurements were done as follows: first, with the field winding alone activated; then, with the armature winding alone activated; finally, with both circuits activated. These measurements made it possible to examine the linearity of the system (i.e., whether the B fields produced by the field and armature windings separately were really additive), and to identify clearly the contributions of each winding to the net field and force in each case. It was important to know whether

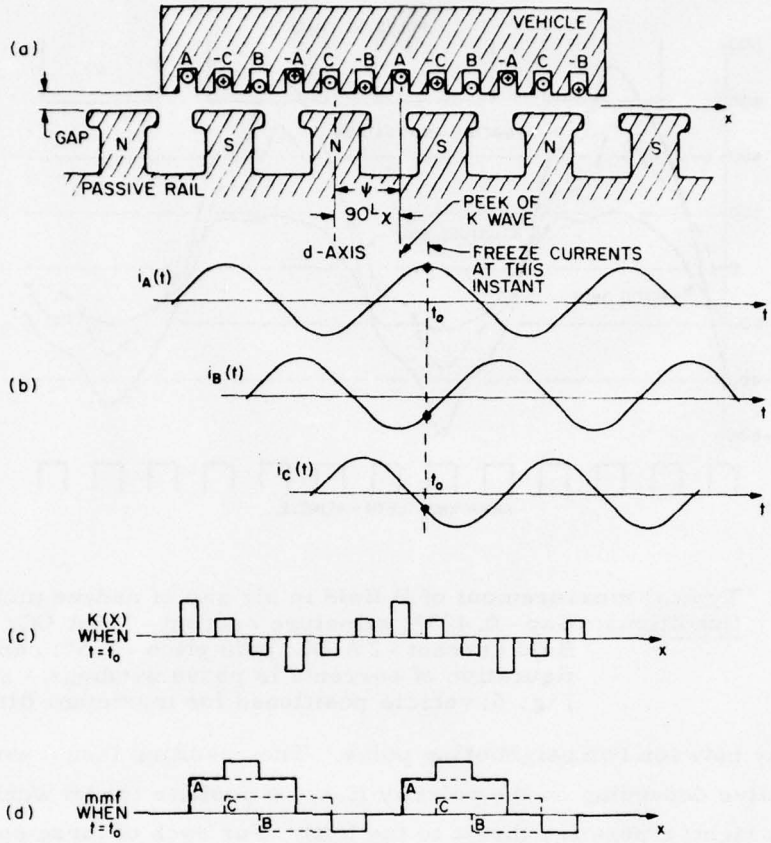


Fig. 5. Sketch illustrating rationale for DC excitation of armature. (a) vehicle and rail structure; (b) currents in armature winding; (c) current distribution along armature; (d) m mf distribution along armature.

linearity was present if a meaningful comparison of the measurement with theory was to be made. Results of a typical B field measurement are shown in Fig. 6, in which near-linearity is apparent.

All force and field measurements were done using 2 A DC for the field winding and 1.5 A DC for the maximum current in the armature winding. This represents a sinusoidal armature current of $1.5/\sqrt{2}$ A r.m.s.

The developed thrust or levitation depends on the relative position of the peak of the magnetic field generated by the armature winding with respect to the poles of the track. When the peak of the armature-generated magnetic field directly faces the center of one of the track poles, as in Fig. 6, it will create either maximum or minimum levitation, depending on polarity. Maximum thrust is achieved when the peak of the magnetic

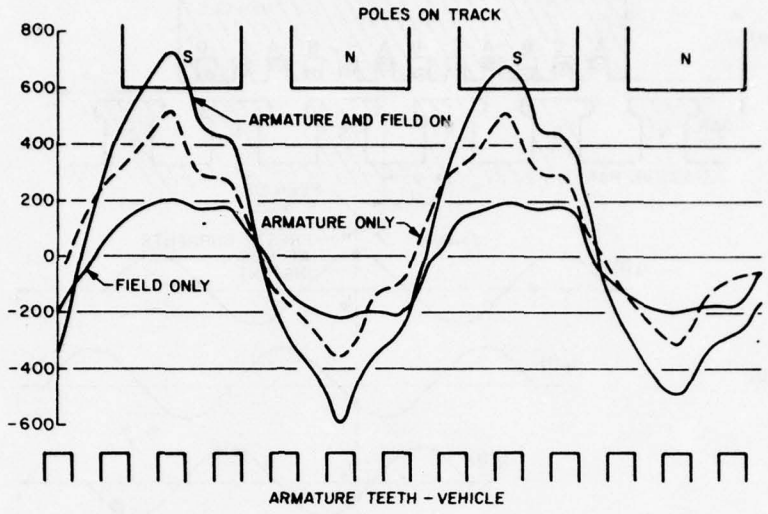


Fig. 6. Typical measurement of B field in air gap of nadyne motor
Conditions: Gap - 0.475"; armature current - 1.5 A DC; field current - 2 A DC; pole pitch - 1.5"; configuration of currents in phase windings - see Fig. 5; vehicle positioned for maximum lift force.

field lies halfway between two neighboring poles. The resulting thrust would be either positive or negative depending on the polarity (i.e., a positive thrust would tend to move the motor to the right; a negative thrust to the left). For each of three configurations representing three different m mf distributions along the armature, field measurements were taken for four different relative positions between the armature and the track: for maximum positive and negative thrust, and for maximum and minimum levitation.

All of the field measurements were taken using the same gap, 0.475 inches, just large enough to permit the insertion of the Hall probe of the gaussmeter. Force measurements were taken for gaps ranging from 3/16 to 6/16 inch, in increments of 1/16 inch.

B. Comparison of Predicted and Measured Forces

Calculation of the thurst and normal forces yields for the 3/8 inch gap:

Thrust - $T = 0.61 \sin \chi - .031 \sin 2\chi$ lbs;

Lift - $F_n = -63.2 + 0.91 \cos \chi - .024 \cos 2\chi$ lbs.

In Fig. 7 is shown a comparison of this prediction with the measured thrust and lift

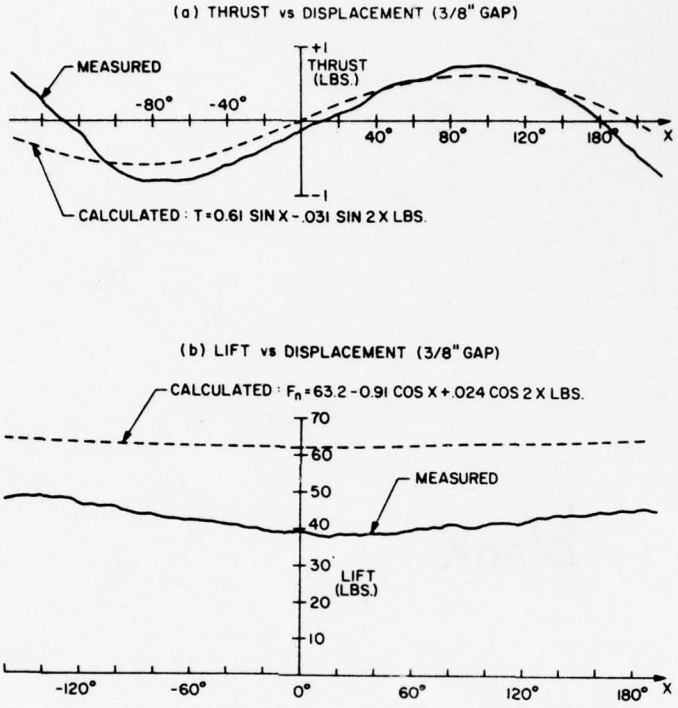


Fig. 7. Comparison of measured and predicted (a) thrust and (b) lift forces in nadyne model.

forces. The degree of correspondence between the two is really a measure of how useful is the theory of the nadyne in describing a real situation. The theory is based on an endless vehicle, and hence does not take into account end effects, absent in a rotating machine; nor does it include detailed consideration of effects of armature teeth or of fringing flux at the pole edges. In addition, there is still present in the model some degree of tilt due to the imperfect rigidity of the track suspension. Although the theory includes intelligent estimates of the leakage flux effects, no precision can be claimed here. When all of these factors are considered, the underestimation of the thrust force (by about 25%) and the overestimation of the lift force (by about 30%) represent, perhaps, reasonably good predictions.

Department of Transportation
DOT-FR-30030

National Science Foundation
EPP75-04656

F. Lalezari, E. Levi and L. Birenbaum

REFERENCE

1. E. Levi, "Preliminary Design Studies on Iron-Cored Synchronously Operating Linear Motors," Final Report No. POLY EE/EP 76-005 under Contract No. DOT-FR-30030 (May 1976). See Part IV, pp. 4-13, 14, 15.

PRECEDING PAGE BLANK-NOT FILMED

II. SYSTEMS

A. COMMUNICATIONS

Program Director: L. Kurz

B. COMPUTERS AND COMPUTER-COMMUNICATION NETWORKS

Program Director: E.J. Smith

C. SAFETY, RELIABILITY AND SOFTWARE ENGINEERING

Program Director: M.L. Shooman

D. SYSTEMS, CONTROL AND NETWORKS

Program Director: D.C. Youla

E. DATA PROCESSING

Program Director: A. Papoulis

THE CLASS OF LEHMANN ALTERNATIVES AS MEANS FOR EVALUATION OF PERFORMANCE IN ROBUST DETECTION

B. Gotz and L. Kurz

Numerous problems in communications, radar, sonar, pattern recognition and picture processing may be posed in terms of hypotheses testing of stochastically ordered hypotheses. The most commonly used model of shift in the mean (change in location) is an example of such ordering. In Ref. 1, Lehmann proposed a new and useful method of realizing an alternative that is stochastically larger than a given hypothesis. For a particular hypothesis distribution $F(x)$ he considers an alternative distribution $G(x) = h(F(x))$, where h is some transformation on $F(x)$ such that $G(x) \leq F(x)$, i.e., the alternative hypothesis is stochastically larger than the null hypothesis. A particularly simple transformation that is proposed in Ref. 1 and that is utilized here is

$$G(x) = F^{1+\theta}(x) \quad \theta > 0 \tag{1}$$

The stochastic ordering property, $G(x) \leq F(x)$, follows from the fact that $F(x)$ is a c.d.f. The class of alternatives specified by Eq. (1) will be referred to as the class of Lehmann alternatives. The Lehmann alternatives have a special analytic convenience for evaluation of performance in robust detection. For instance, the conceptually useful efficacy (local rate of change in signal-to-noise ratio) becomes for many detectors independent of the particular reference distribution $F(x)$. There remains the major question of relevance. In this connection it should be pointed out that in distribution-free detection procedures the precise nature of the hypothesis separation will not be likely known. "What is then required are alternatives representative of principal types of deviation from the hypothesis, in terms of which one can study, at least in outline, the ability of various tests to detect such deviations" (p. 24 of Reference 1). The effort here will be concentrated on how Lehmann alternatives compare to other kinds of hypothesis separations and on the nature of the efficacy calculations for these alternatives in m-interval detectors.[2]

A. Lehmann Alternatives vs. Other Kinds of Hypothesis Separation

The question is: how does one relate the Lehmann indexing parameter θ in

$$H_o : F(x) \qquad H_1 : F^{1+\theta}(x)$$

to other indexing parameters?

The change in the mean and variance is a natural measure of hypothesis separation. Under the hypothesis with the c.d.f. $F(x)$ and p.d.f. $f(x)$

$$H_o: \; E\left[x/H_o\right] = \int_{-\infty}^{\infty} x\, f(x)\, dx$$

$$\sigma^2_{x/H_o} = E\left[\left(x - E(x/H_o)\right)^2\right]$$

and

$$H_1: F^{1+\theta}(x) = G(x), \; g(x) = (1+\theta)\, F^{\theta}(x)\, f(x), \; \theta > 0$$

the change in the mean is

$$\Delta m = \int_{-\infty}^{\infty} x \left[(1+\theta)\, F^{\theta}(x) - 1\right] f(x)\, dx$$

and the change in variance is

$$\Delta\sigma^2 = \int_{-\infty}^{\infty} \left[x - \left(\Delta m + E(x/H_o)\right)\right]^2 (1+\theta)\, F^{\theta}(x)\, f(x)\, dx - \sigma^2_{x/H_o}$$

For a specific distribution the evaluation of the above integrals as a function of θ is generally difficult and one must resort to numerical methods. In Fig. 1 results of

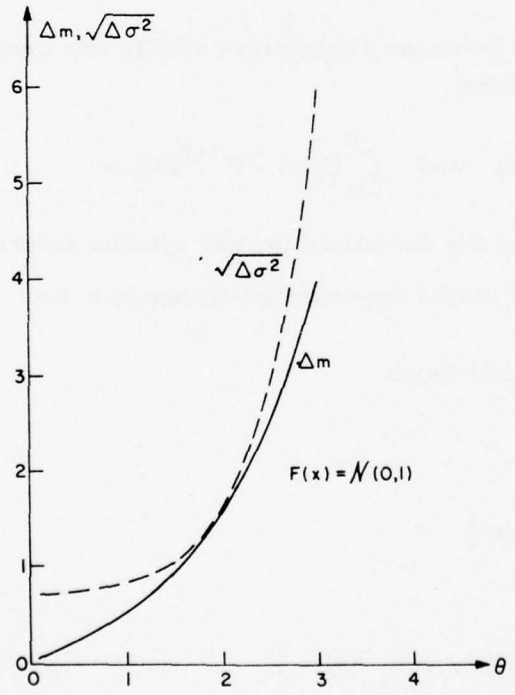


Fig. 1. Change in mean and standard deviation accompanying a Lehmann alternative $F^{1+\theta}(x)$ where $F(x) = N(0, 1)$.

numerical calculations based on a three-point Gaussian-Hermite quadrature formula for the reference c.d.f. N(0,1) are shown. Similar calculations are readily carried out for other reference distributions.

A more compact comparison that measures hypotheses separation is the unweighted change in area between the hypothesis and alternative c.d.f., i.e.,

$$d(\theta) = \int_{-\infty}^{\infty} [F(x) - F_\theta(x)]\, dx$$

Since

$$d(\theta) < \int_{-\infty}^{\infty} F(x)\, dx$$

and if

$$\lim_{\theta \to 0} [F(x) - F_\theta(x)] = F(x)$$

as is true for any physically meaningful indexing, it follows that

$$\lim_{\theta \to 0} d(\theta) = \int_{-\infty}^{\infty} F(x)\, dx$$

Both shift of the mean and Lehmann alternatives satisfy this condition with the separation expressions, $d(\theta)$, of the form

$$\int_{-\infty}^{\infty} [F(x) - F(x-\theta)]\, dx \quad \text{and} \quad \int_{-\infty}^{\infty} [F(x) - F^{1+\theta}(x)]\, dx$$

respectively. Both are readily calculated for any specific reference c.d.f. $F(x)$.

A useful variation of $d(\theta)$ is the weighted change in c.d.f.

$$\tilde{d}(\theta) = \int_{-\infty}^{\infty} [F(x) - F_\theta(x)]\, f(x)\, dx$$

Since

$$\tilde{d}(\theta) < \int_{-\infty}^{\infty} F(x)\, f(x)\, dx = \frac{1}{2}$$

and if

$$\lim_{\theta \to 0} [F(x) - F_\theta(x)] = F(x), \quad \lim_{\theta \to 0} \tilde{d}(\theta) = \frac{1}{2} \quad .$$

The special convenience of the separation measure, $\tilde{d}(\theta)$, is the fact that for Lehmann alternatives of the type $F_\theta(x) = F^{1+\theta}(x)$

$$\tilde{d}(\theta) = \int_{-\infty}^{\infty} [F(x) - F^{1+\theta}(x)]\, f(x)\, dx = \int_0^1 (y - y^{1+\theta})\, dy = \frac{1}{2} - \frac{1}{2+\theta}$$

independent of the reference c.d.f. $F(x)$. For another type of hypothesis indexing and some reference c.d.f., $\tilde{d}(\theta)$ need be calculated and compared to $\frac{1}{2} - \frac{1}{2+\theta}$.

A comparison of $\tilde{d}(\theta)$ for shift alternatives and standardized Cauchy and Gaussian c.d.f.'s to Lehmann alternatives is shown in Figure 2. Calculations were performed using a 4-point Gaussian-Hermite quadrature formula. The results of Fig. 2 imply that $\tilde{d}(\theta)$ for Lehmann alternatives represents a conservative bound on the separation measure for shift alternatives.

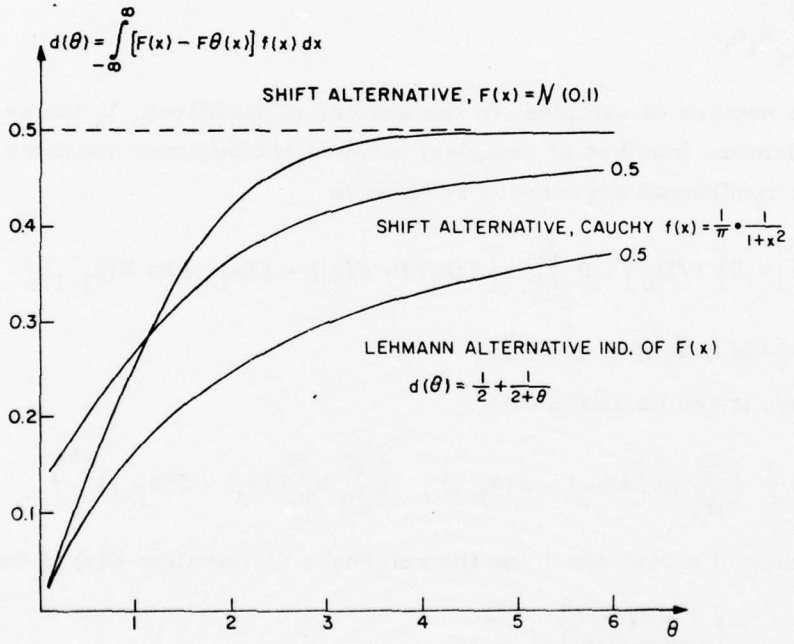


Fig. 2. A hypothesis separation measure compared for various alternatives.

B. Lehmann Alternative Efficacy Calculations

Let

$H_o : F_o(x) = F(x)$

$$H_1 : F_1(x) = F^{1+\theta}(x) = e^{(1+\theta)\,\ell n\, F(x)} = g(\theta) = F_\theta(x),\ \theta > 0$$

By the mean value theorem

$$g(\theta) = g(0) + \theta g'(\hat{\theta}) \qquad 0 < \hat{\theta} < \theta$$

and, assuming that $g'(\theta)$ is continuous in θ, for small θ

$$g(\theta) \doteq g(0) + \theta g'(0)$$

where $g(0) = F(x)$ and

$$g'(0) = [\ell n\, F(x)]\, e^{(1+\theta)\,\ell n\, F(x)}\Big|_{\theta=0} = F(x)\, \ell n\, F(x)$$

For small θ

$$F_\theta(x) \doteq F(x) + \theta\, F(x)\, \ell n\, F(x) \tag{2}$$

For m-interval detectors, the test statistic for efficacy calculations is

$$T = \frac{1}{L} \sum_{i=1}^{m} b_i n_i$$

where L is the number of samples, m the number of partitions, b_i the scores and n_i are occupancy indicators (number of samples) between two adjacent quantiles. It can be shown that the conditional expectation reduces to

$$E\left[T/H_1\right] = E\left[T/H_o\right] + \theta \sum_{j=1}^{m} b_j \left[F(a_j)\, \ell n\, F(a_j) - F(a_{j-1})\, \ell n\, F(a_{j-1})\right] \tag{3}$$

where $\{a_j\}$ are the quantiles of $F(x)$.

Similarly, it can be shown that

$$\sigma^2_{T/H_o} = \frac{1}{L} \left\{ \sum_{j=1}^{m} b_j^2 \left[F(a_j) - F(a_{j-1})\right] - \left[\sum_{j=1}^{m} b_j \left[F(a_j) - F(a_{j-1})\right]\right]^2 \right\} \tag{4}$$

Since the efficacy of a detector T for the reference distribution $F(x)$ is defined by

$$\mathcal{E}_{T,F} = \lim_{L\to\infty} \frac{\left[\dfrac{\partial E(T/H_1)}{\partial \theta}\Big|_{\theta=0}\right]^2}{L\, \sigma^2_{T/H_o}} \tag{5}$$

it follows from Eqs. (3-5) that

$$\mathcal{E}_{T,F} = \frac{\left\{\sum_{j=1}^{m} b_j\left[F(a_j)\,\ell n\,F(a_j) - F(a_{j-1})\,\ell n\,F(a_{j-1})\right]\right\}^2}{\sum_{j=1}^{m} b_j^2\left[F(a_j) - F(a_{j-1})\right] - \left\{\sum_{j=1}^{m} b_j\left[F(a_j) - F(a_{j-1})\right]\right\}^2} \tag{6}$$

Of particular interest is the special case of equiprobable partitioning with $F(a_j) - F(a_{j-1}) = \frac{1}{m}$ and $F(a_j) = (\frac{1}{m})j$, $j = 0, 1, \cdots, m$. For equiprobable partitioning, Eq. (6) reduces to

$$\mathcal{E}_{T,F} = \frac{\left\{\sum_{j=1}^{m} b_j[j\,\ell n\,j - (j-1)\,\ell n(j-1) - \ell n\,m]\right\}^2}{m\sum_{j=1}^{m} b_j^2 - (\sum_{j=1}^{m} b_j)^2} \tag{7}$$

As was shown in Refs. 2 and 3, for the optimal m-interval detector

$$b_i = \ell n\left\{m\left[F_1(a_i) - F_1(a_{i-1})\right]\right\}$$

which reduces for small signal-to-noise ratio (i.e., locally) to

$$b_i = \theta[i\,\ell n\,i - (i-1)\,\ell n(i-1) - \ell n\,m] \tag{8}$$

where use has been made of the approximation $\ell n(1+\xi) \doteq \xi$ for small ξ. From Eqs. (7) and (8), it follows that for Lehmann alternatives and equiprobable partitioning, the efficacy reduces to

$$\mathcal{E}_{T,F} = \frac{1}{m}\sum_{j=1}^{m}(j\,\ell n\,j - (j-1)\,\ell n(j-1) - \ell n\,m)^2 \tag{9}$$

which is independent of the reference c.d.f. F(x).

After some mathematical manipulations, it can be shown that the limiting Lehmann alternative efficacy expression for the locally most powerful m-interval detector is

$$\lim_{m\to\infty} \mathcal{E}_{T,F} = 1 \tag{10}$$

The result of Eq. (10) must agree with the general local efficacy bound

$$\mathcal{E}_{T,F} \le \text{Inf}\ F_{\theta=0} = E\left\{\left[\frac{\partial}{\partial\theta}\,\ell n\,f_\theta(x)\right]^2\Big|_{\theta=0}\right\} \tag{11}$$

In addition, as for the shift alternative, it is expected that as $m\to\infty$, the efficacy of the locally most powerful m-interval detector approaches the efficacy of the locally most powerful detector on the raw data, i.e., $\lim_{m\to\infty} E_{T,F} = \text{Inf}\ F_{\theta=0}$, which would require

$$\text{Inf } F_{\theta=0} = 1$$

independently of F(x). This indeed is the case because

$$F_\theta(x) = F^{1+\theta}(x) \rightarrow f_\theta(x) = (1+\theta)\, F^\theta(x)\, f(x)$$

$$\rightarrow \ell n\, f_\theta(x) = \ell n(1+\theta) + \theta\, \ell n\, F(x) + \ell n\, f(x)$$

and

$$\text{Inf } F_{\theta=0} = E\left\{\left[\frac{\partial}{\partial\theta}\, \ell n\, f_\theta(x)\right]^2_{\theta=0}\right\} = E\left\{\left[1 + \ell n\, F(x)\right]^2\right\} = 1$$

Thus, it can be concluded that for Lehmann alternatives

$$E_{T,F} \leq 1$$

It should be indicated at this point that in comparing detectors using asymptotic relative efficiency by comparing efficacies, one must restrict oneself to a specified alternative hypothesis indexing. Comparing efficacies with different alternative indexing is meaningless.

Joint Services Technical Advisory Committee
F44620-74-C-0056

B. Gotz and L. Kurz

REFERENCES

1. E. L. Lehmann, "The Power of Rank Tests," Ann. Math. Stat., 24, 23-43 (March 1953).

2. Y. C. Ching and L. Kurz, "Nonparametric Detection Based on m-Interval Partitioning," IEEE Trans. Information Theory, IT-17, 251-257 (March 1972).

3. L. Kurz, "Nonparametric Detectors Based on Partition Tests," Supplementary Notes, EE 704, Polytech. Inst. of New York (September 1976).

SIGNAL DESIGN AND RANK M-ARY DETECTION

Y.C. Ching and L. Kurz

In this report, M-ary nonparametric detectors are considered with the associated signalling alphabets. Unlike the class of parametric detectors for which the classification procedures are based on the optimum partitioning of the sample space associated with the received signals corrupted by noise of known distribution, the nonparametric detectors are designed by first selecting a nonparametric partitioning of the sample space. In turn, the latter selection, motivated primarily by the type of limited knowledge available about the noise distribution and by the simplicity and flexibility of the resulting detector structure, governs the selection of signalling alphabets. This type of systems are particularly useful in a severe noise environment and, because of their M-ary nature, these systems utilize the available bandwidth efficiently. Though M-ary nonparametric detectors were considered previously,[1,2] in this report the problem of joint detector and signal selection based on nonparametric partitioning of the space is considered for the first time. In particular, the nonparametric detectors which are based on the rank partitioning (rank detectors) are considered. For the rank detectors, a set of signal vectors which are permutations of one another is selected.

Two rank detectors, both using Chernoff-Savage[3] test statistics, are proposed for the signalling alphabet. Due mainly to the asymptotic normality of the Chernoff-Savage test statistics, independent of which signal has been transmitted, the resulting detectors are equivalent to a maximum likelihood detector based on the test statistics. The performance of these detectors, as measured by Pitman-Noether ARE, is shown to be the same as their binary counterparts.

A. Signal Selection for M-ary Nonparametric Detection

Denote by $U = \{u^{(i)}, i = 1, \cdots, M\}$ the set of possible signal vectors and $V = \{v_j, j = 1, \cdots, N\}$ the set of nonparametric partitions and let $u^{(r)}$, $Z^{(r)} \in v_r$ and $u^{(s)} \in v_s$. Denote by $d(x,y)$ the distance between x and y, where x and y could be either vectors or nonparametric partitions, or a mixture of both. Because the exact noise distribution is unknown, the choice of a distance is quite arbitrary. However, by definition of nonparametric partitioning, it is assumed that

$$d(u^{(s)}, u^{(r)}) = d(u^{(s)}, u^{(r)}) \tag{1}$$

$$= d(u^{(s)}, v_r) \tag{2}$$

$$= d(v_s, v_r) \tag{3}$$

except for the vectors on the boundaries of the nonparametric partitions. The detection

rule is to assume that $u^{(r)}$ was transmitted, for a received observable Z, if

$$d(Z, u^{(r)}) < d(Z, u^{(i)}) \quad \text{for all} \quad i \neq r \tag{4}$$

However, if there are two signal vectors $u^{(r)}$ and $u^{(s)}$ belonging to the same nonparametric partition, v_r, then

$$d(Z, u^{(r)}) = d(Z, v_r) \tag{5}$$

$$= d(Z, u^{(s)}) \quad \text{for all} \quad Z \tag{6}$$

and there is no way to distinguish $u^{(r)}$ and $u^{(s)}$ unless the noise distribution is known, in which case the problem reduces to a parametric one. From these considerations, it is clear that the signal selection for a nonparametric M-ary detection problem depends on the manner of nonparametric partitioning. Specifically, there can be at most one signal vector per nonparametric partition or $M \leq N$.

In this report, only M-ary rank detectors are considered. The rank detectors are based on the nonparametric partitions which result from separating the n-dimensional Euclidean sample space by the set of hyperplanes

$$\{z_i = z_j, \text{ all } i \neq j\} \tag{7}$$

The set of rank partitions $\{R_j, j = 1, 2, \cdots, n!\}$ is isomorphic to the set of rank vectors $\{R^{(j)}, j = 1, \cdots, n!\}$. Furthermore, since each rank vector is contained in a rank partition, it can be regarded as a representative vector of the rank partition. Often the distance between the rank partitions is defined on the basis of their representative vectors. For example, Spearman's rank correlation coefficient is defined as the Euclidean distance between the rank vectors.

For a rank partition R_i, if $u^{(i)} \in R_i$, then

$$a\,u^{(i)} \in R_i \tag{8}$$

for any positive constant a. Thus

$$d(Z, u^{(i)}) = d(Z, R_i) \tag{9}$$

$$= d(Z, a\,u^{(i)}) \tag{10}$$

for all received observables Z. It is, therefore, logical to limit the signals to the set of equal-energy signals, i.e., if

$$u^{(i)} = (X_1^{(i)} X_2^{(i)} \cdots X_n^{(i)}) \tag{11}$$

then

$$E^{(i)} = \sum_{k=1}^{n} (X_k^{(i)})^2 = E, \quad i = 1, 2, \cdots, M \tag{12}$$

where E represents the energy associated with the signalling set U.

The actual signalling set is selected as follows. Let M be the number of signals and n be the length of the signal vector. For simplicity, assume that b and q are some integers such that

$$n = bq \tag{13}$$

$$M < b! \tag{14}$$

Let

$$u^{(1)} = (\tilde{a}_1 \tilde{a}_2 \cdots \tilde{a}_b) \tag{15}$$

be one of the signals, where

$$\tilde{a}_k = a_k(1\ 1 \cdots 1) \tag{16}$$

is a vector of length q and

$$a_1 < a_2 < \cdots < a_b \tag{17}$$

The signalling set is then formed by permuting $\tilde{a}_k$ s. It should be noted that the permutation is performed on q components of $u^{(1)}$ as an entity and a_k corresponds to the k-th level of the signal $u^{(1)}$. Each level is repeated q times to increase the reliability and also to guarantee that the test statistics would be approximately normally distributed. This normality is necessary to construct a maximum likelihood detector. There are at most b! signal vectors and each can be represented by vectors of length b of the form

$$u^{(i)} : S^{(i)} = (s_1^{(i)}\ s_2^{(i)} \cdots s_b^{(i)}) \tag{18}$$

which are permutations of

$$S^{(1)} = (a_1 a_2 \cdots a_b) \tag{19}$$

To each signal $S^{(i)}$ corresponds a rank vector of length b, $R_b^{(i)}$, which is the appropriate permutation of the identity rank vector

$$R_b^{(1)} = (12 \cdots b) \tag{20}$$

At the receiver, the observable is a vector of length bq of the form

$$Z = (z_{11} \cdots z_{1q}, z_{21} \cdots z_{2q}, \cdots, z_{b1} \cdots z_{bq}) \tag{21}$$

Let the noise vector be denoted by

$$N = (n_{11} \cdots n_{1q}, \cdots, n_{b1} \cdots n_{bq}) \tag{22}$$

then, when $u^{(i)}$ is transmitted,

$$z_{kr} = s_k^{(i)} + n_{kr} \qquad \begin{matrix} k = 1, 2, \cdots, b \\ r = 1, 2, \cdots, q \\ i = 1, 2, \cdots, M \end{matrix} \tag{23}$$

The noise process is assumed to be stationary and the sampling rate sufficiently low so that the n_{kr}'s are i.i.d. random variables with an unknown distribution $F(z)$. Denote

$$Z_k = \{z_{kr}, r = 1, 2, \cdots, q\} \tag{24}$$

as the k-th sample of the observable Z and the corresponding rank vector by

$$R(Z) = (r_1 r_2 \cdots r_b) \tag{25}$$

where r_j is the rank of the j-th sample Z_j among $\{Z_k, k = 1, \cdots, b\}$. Since

$$Z_k : F(z - s_k^{(i)}) \tag{26}$$

it follows that, when $u^{(i)}$ is transmitted,

$$R(Z) = R_b^{(i)} \tag{27}$$

Thus a detector which senses the stochastic ordering of all Z_k, $k = 1, 2, \cdots, b$ can differentiate among transmitted signals.

B. Two Classes of Detectors

The detectors proposed here are motivated by the form of the optimum M-ary Gaussian noise detector. As for the Gaussian noise detector, the detector is decomposed into two parts: a sample processor and M matching networks, each corresponding to an element of the signalling set. However, while only the sample mean (or Student's t test) is generated in the sample processor of the Gaussian noise detector, nonparametric test statistics are formed in the sample processor of the proposed detectors. The matching networks of the nonparametric detectors are designed as follows. Let L_j be the output of the matching network corresponding to a signal $u^{(j)} \in U$, and $E[L_i/u^{(j)}]$ be the conditional expectation of L_i when $u^{(j)}$ is transmitted.

The matching network is a device with the property

$$E[L_j/u^{(j)}] = \max_i \{E[L_i/u^{(j)}]\} \tag{28}$$

while some optimality criterion is satisfied. Since in most of the cases the nonparametric tests tend to be singular for large sample sizes, the matching networks ensure the consistency of the detector. Finally, the detector compares all L_j's and selects the signal $u^{(j)}$ corresponding to $L_j = \max_i \{L_i\}$.

1. Detector I

In this detector the sample processor forms the k-sample version of the two-sample Chernoff-Savage test statistics as developed by Puri.[4] If the bq observations of vector Z are ordered so that

$$z_{(1)} < z_{(2)} < \cdots < z_{(bq)} \tag{29}$$

the Puri T_k-statistic is defined by

$$T_k = \frac{1}{q} \sum_{r=1}^{bq} d_r C_{rk} \qquad k = 1, \cdots, b \tag{30}$$

where d_r's are ordered constants satisfying certain conditions[4] and

$$C_{rk} = \begin{cases} 1 & \text{if } z_{(r)} \text{ is from } Z_k \\ 0 & \text{otherwise} \end{cases} \tag{31}$$

All b of the T_k-statistics are shaped by the appropriate matching networks, yielding a new test statistic

$$L_i = \sum_{k=1}^{b} T_k s_k^{(i)} \tag{32}$$

Since

$$\sum_{k=1}^{b} T_k = \frac{1}{q} \sum_{r=1}^{bq} d_r \tag{33}$$

must be satisfied, all T_k's are correlated. It can be shown that the proposed two-stage nonparametric M-ary detector is equivalent to a locally maximum likelihood detector using T_k-statistics.

Introducing

$$H_q(z) = \frac{1}{b} \sum_{k=1}^{b} F_{qk}(z) \tag{34}$$

where $F_{qk}(z)$ is the empirical distribution of Z_k and $H_q(z)$ is the empirical distribution of Z and defining the score-generating function

$$J(x) = \lim_{q \to \infty} J_q(x) \tag{35}$$

where

$$J_q(r/bq) = dr \tag{36}$$

the T_k-statistic may be rewritten as

$$T_k = \int_{-\infty}^{\infty} J_q[H_q(z)]\, dF_{qk}(z), \quad k = 1, 2, \cdots, b \tag{37}$$

It can be shown that T_k's are asymptotically normal. Furthermore, defining the vectors

$$T = (T_1 T_2 \cdots T_b)' \tag{38}$$

and

$$\mu = (\mu_1 \mu_2 \cdots \mu_b)' \tag{39}$$

where

$$\mu_k = \int_{-\infty}^{\infty} J(H(z))\, dF_k(z), \quad k = 1, 2, \cdots, b \tag{40}$$

then the vector statistic $q^{1/2}(T - \mu)$ has a limiting normal distribution with zero-mean and known covariance function. For each of the signals

$$E[T/u^{(j)}] = \mu^{(j)}$$

and

$$\text{cov}(T/u^{(j)}) = \sum_j$$

The matching networks of the detector are formed by

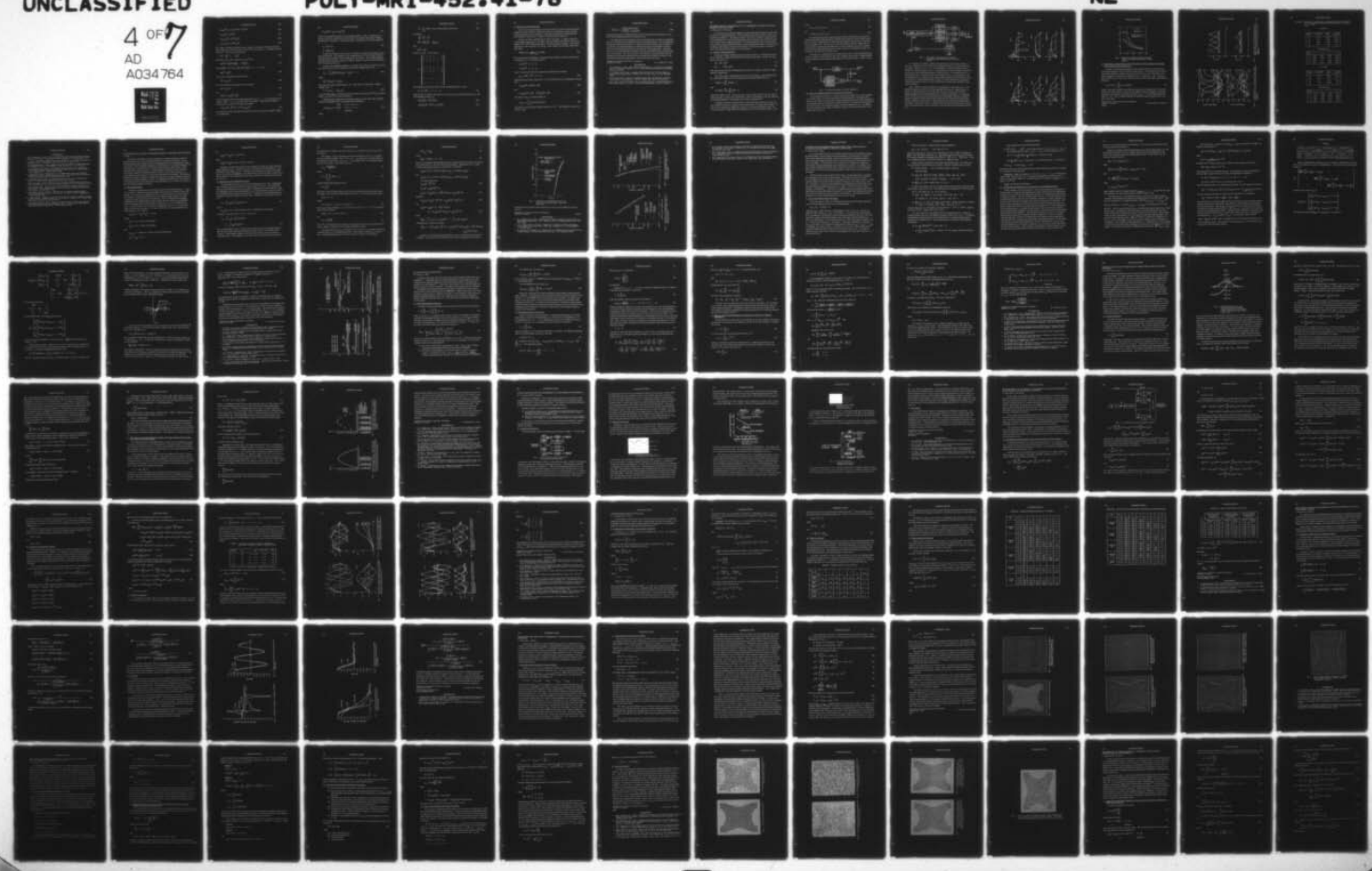
$$L_i = T' S^{(i)}, \quad i = 1, 2, \cdots, M \tag{41}$$

Since T is asymptotically normal so are L_i's (linear transformations), or L_i's are completely specified asymptotically by

AD-A034 764 POLYTECHNIC INST OF NEW YORK BROOKLYN MICROWAVE RESE--ETC F/G 20/3
PROGRESS REPORT NUMBER 41, 15 SEPTEMBER 1975 THROUGH 14 SEPTEMB--ETC(U)
NOV 76 A A OLINER F44620-74-C-0056

UNCLASSIFIED POLY-MRI-452.41-76 NL

4 OF 7
AD
A034764



$$E[L_i/u^{(i)}] = E[T'/u^{(i)}]S^{(i)} = \mu^{(i)}S^{(i)} \tag{42}$$

$$E[L_i/u^{(j)}] = \mu^{(j)}S^{(i)} \tag{43}$$

$$\text{var}[L_i/u^{(i)}] = S^{(i)'}\sum_i S^{(i)} \tag{44}$$

$$\text{var}[L_i/u^{(j)}] = S^{(i)'}\sum_j S^{(i)} \tag{45}$$

Since J(x) is usually a monotone function, there is a strong correlation between the transmitted signal $u^{(j)}$ and the mean vector $\mu^{(j)}$ of the statistic T. Consider a specific signal $u^{(1)}$. It can be easily shown that

$$\mu_1^{(1)} < \mu_2^{(1)} < \cdots < \mu_b^{(1)} \tag{46}$$

Denote the rank vector of the mean vector $\mu^{(i)}$ by

$$R(\mu^{(i)}) = \left(R(\mu_1^{(i)})\ R(\mu_2^{(i)})\ \cdots\ R(\mu_b^{(i)})\right)' \tag{47}$$

where $R(\mu_r^{(i)})$ *is the rank of* $\mu_r^{(i)}$ *among* $\{\mu_k^{(i)} : k = 1, \cdots, b\}$, *then*

$$R(\mu^{(1)}) = R_b^{(1)} \tag{48}$$

Let Q_{1i} be a permutation matrix such that

$$S^{(i)} = Q_{1i}\,S^{(1)} \tag{49}$$

Since the signals are permutations of one another

$$\mu^{(i)} = Q_{1i}\,\mu^{(1)} \tag{50}$$

and

$$R(\mu^{(i)}) = Q_{1i}R_b^{(1)} = R_b^{(i)} \tag{51}$$

or the stochastic order of the random variables $\{T_k : k = 1, \cdots, b\}$ is the same as the order of $\{s_k^{(i)} : k = 1, 2, \cdots, b\}$ when $u^{(i)}$ is transmitted. Forming the conditional expectation of L_i when $u^{(j)}$ has been transmitted,

$$E[L_i/u^{(j)}] = \mu^{(j)'}S^{(i)} = \mu^{(1)'}Q_{1j}'Q_{1i}S^{(1)} \tag{52}$$

and using the fact that the right-hand side of Eq. (52) is maximized if $R(\mu^{(j)}) = R_b^{(i)}$, it is concluded that

$$E[L_j/u^{(j)}] = \max_i \{E[L_i/u^{(j)}\} \tag{53}$$

which is the desirable property of the matching networks. Since, in addition, the covariance matrix $\sum_j$ approaches the null matrix as the sample size tends to infinity, this detector is consistent. The decision rule is, therefore, to accept the signal $u^{(j)}$ if

$$L_j = \max_i \{L_i\} \tag{54}$$

2. Detector II

In practical systems it may be more convenient to compare the samples Z_k pairwise and form $b(b-1)/2$ statistics in the sample processor. Each of these statistics will indicate the stochastic order between a pair of samples. The matching networks will perform the selection of signals as before. However, these networks should have $b(b-1)/2$ inputs instead of b.

For the same signalling alphabet as in Section A and a pair of samples Z_r and Z_s, the statistics formed in the sample processor of the modified detector are

$$T_{rs} = \int_{-\infty}^{\infty} J_q(H_q^{rs}(z))\, d(F_{qr}(z) - F_{qs}(z)) \qquad r < s \tag{55}$$

where

$$H_q^{rs} = \frac{1}{2}(F_{qr}(z) - F_{qs}(z))$$

and $F_{qr}(z)$ and $F_{qs}(z)$ are the sample c.d.f. of Z_r and Z_s, respectively. Define a random vector

$$V = (T_{12}\, T_{13} \cdots T_{(b-1)b})' \tag{56}$$

it can be shown that V is asymptotically normal with a known mean $\eta^{(i)}$ and covariance matrix $\Lambda_{(i)}$ when $u^{(i)}$ has been transmitted.

When the signal separation is small (low signal-to-noise ratio), the covariance matrix is independent of the transmitted signal resulting in

$$\operatorname{cov}(T_{ki}, T_{rs}) = \begin{cases} A^2/q & \text{if } k = r \text{ or } i = s \\ -A^2/q & \text{if } k = s \text{ or } i = r \\ 0 & \text{otherwise} \end{cases} \tag{57}$$

where

$$A^2 = 2 \iint_{x<y} F(x)(1 - F(y))\, J'(F(x))\, J'(F(y))\, dF(x)\, dF(y) \tag{58}$$

Introducing

$$s_{rs}^{(i)} = s_r^{(i)} - s_s^{(i)}$$

$$\underset{\sim}{S}^{(i)} = (s_{12}^{(i)}\ s_{13}^{(i)} \cdots s_{(b-1)b}^{(i)})'$$

then

$$\underset{\sim}{S}^{(i)} = G S^{(i)} \tag{59}$$

where G is a b(b - 1) by b matrix

$$G = \begin{bmatrix} 1 & -1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & -1 & 0 & \cdots & 0 \\ 1 & 0 & 0 & -1 & \cdots & 0 \\ & & & & \ddots & \\ 1 & 0 & 0 & 0 & \cdots & -1 \\ 0 & 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & 0 & -1 & \cdots & 0 \\ & & & & \ddots & \\ 0 & 1 & 0 & 0 & \cdots & -1 \\ & & \cdots & & & \\ 0 & \cdots & & 1 & & -1 \end{bmatrix} \tag{60}$$

The modified test statistics formed in the matching networks are then

$$W_i = \frac{1}{b} V' \underset{\sim}{S}^{(i)} \qquad i = 1, 2, \cdots, M \tag{61}$$

Since all W_i's are linear combinations of V, they are also asymptotically normal. The corresponding moments are

$$E[W_i/u^{(j)}] = \eta^{(j)} G S^{(i)}/b \tag{62}$$

$$\mathrm{Var}[W_i/u^{(j)}] = S^{(i)'} G' \Lambda_j G S^{(i)}/b^2 \tag{63}$$

C. Performance of the Detectors

In this section the performance of the two detectors in terms of Pitman-Noether asymptotic relative efficiency (ARE) with respect to the optimum detector (sample mean) in gaussian noise is considered. Since the two detectors are equivalent for small signal-to-noise ratios, it suffices to find the ARE for detector I.

Usually, the ARE is defined only for binary detectors. In order to adapt the criterion to M-ary detectors, we define a pair-wise ARE of two detectors used to test whether $u^{(i)}$ or $u^{(j)}$ has been transmitted. Denote this ARE as ARE(ij). The relative performance of two detectors, D_1 and D_2, is then rated by the average of these pair-wise ARE's

$$ARE(D_1, D_2) = \frac{2}{M(M-1)} \sum_{i<j} ARE(ij) \tag{64}$$

Since the a priori probabilities of transmission are unknown, the pair-wise detection is symmetric, i.e., the decision rule is

$$\text{accept } u^{(i)} \quad \text{if} \quad L_i > L_j$$

$$\text{accept } u^{(j)} \quad \text{if} \quad L_i < L_j$$

Thus, we can use for the sake of analysis an equivalent test statistic

$$L_{ij} = T'(S^{(i)} - S^{(j)}) \quad \text{all} \quad i \neq j \tag{65}$$

Following the procedure of Section B, it can be shown that L_{ij} is asymptotically normal with

$$E[L_{ij}/u^{(j)}] = \mu^{(j)}(S^{(i)} - S^{(j)}) \tag{66}$$

and

$$\text{var}[L_{ij}/u^{(j)}] = (S^{(i)} - S^{(j)})' \sum_j (S^{(i)} - S^{(j)}) \tag{67}$$

The efficacy of L_{ij} is readily found to be

$$\mathcal{E}(L_{ij}) = \left(\int_{-\infty}^{\infty} \frac{d}{dz} J(F(z)\, dF(z))\right)^2 / A^2 \tag{68}$$

The efficacy of the optimum Gaussian detector is $1/\sigma^2$. The ARE(ij) for any pair of signals is then

$$ARE(i,j) = 2\,\frac{\left(\int \frac{d}{dz} J(F(z))\, dF(z)\right)^2}{\int J^2(x)\,dx - \left(\int J(x)\,dx\right)^2} \tag{69}$$

which is the same as for binary detectors where Chernoff-Savage statistics are compared with the optimum detector in Gaussian noise.

In using one of the two proposed classes of detectors the structure of a particular detector is dictated by the underlying class of noise distributions, ease of implementation and the signal processing time. Though the analytical results based on ARE are inconclusive, the proposed approach is sufficiently general and flexible to permit a choice of a particularly good detector structure for a given application. The latter choice can be easily obtained from digital computer simulations with final adjustments in an actual system. Because of the form of the detectors, the proposed M-ary nonparametric detection and signal selection procedures have similar advantages, when compared with the binary procedures, as in the Gaussian noise case.

Joint Services Technical Advisory Committee
F44620-74-C-0056

Y.C. Ching and L. Kurz

REFERENCES

1. E. D. Soll and L. Kurz, "Most Powerful Rank Detection - An Extension and Application of Hoeffding's Theorem," Proc. 1967 Princeton Conference on Information Sciences, 330-334.
2. J.B. Millard and L. Kurz, "Adaptive Threshold Detection of M-ary Signals in Statistically Undefined Noise," IEEE Trans. on Com. Tech., 601-611 (October 1966).
3. M. Chernoff and I.R. Savage, "Asymptotic Normality and Efficiency of Certain Nonparametric Test Statistics," Annals of Math. Stat., 24, 972-994 (1958).
4. M.L. Puri, "Some Distribution-free k-sample Rank Tests of Homogeneity Against Ordered Alternatives," Comm. on Pure and Applied Math., 18, 51-63 (1965).

ON IMPROVEMENT IN PERFORMANCE OF TRANSMISSION SYSTEMS IN ADDITIVE GAUSSIAN AND BURST NOISE

L. Kurz and R.J. Nawrocky

In recent years, the general problem of signal and detector design in burst noise has been of considerable interest. Several widely varying approaches have been used to attack the problem.[1-9] Recently the authors introduced a method for signal and detector optimization in burst and Gaussian noise using two approaches within the framework of the discrete maximum principle.[10] In this presentation alternate approaches to the problem using multi-threshold and varying threshold procedures, are discussed. In addition, the influence of burst noise on M-ary system performance in Gaussian noise and intersymbol interference are investigated.

A. Variable Threshold Systems

Before proceeding with the description of this procedure, consider the decision process based on a single threshold level in Gaussian noise alone. In this case, the input to the detector is

$$H_0: \quad z(k) = n_g(k) \tag{1}$$

$$H_1: \quad = x(k) + n_g(k) \tag{2}$$

where H_0 and H_1 represent the hypotheses that there is no signal or there is a signal present, respectively.

For a zero-mean, white Gaussian noise of variance $N_0/2$, under the hypothesis H_1, the detector output at $k = K$ is Gaussian with mean and variance given by

$$E[y(K)] = \Delta \sum_{k=0}^{K} z(k)\, g(k) = Y_1 \tag{3}$$

and

$$\mathrm{Var}[y(K)] = \frac{N_0}{2} \Delta \sum_{k=0}^{K} g^2(k) = V_1 \tag{4}$$

where the notation of Ref. 10 has been used. It is well known that in this case the optimum detector is based on the comparison of y(K) to a threshold y_0, and the setting of the threshold depends on the preselected criterion of performance.

In additive Gaussian and burst noise, on the other hand, the mean and the variance of y(K) are functions of the burst noise parameters c (location) and q (width). If the receiver nonlinearity suppresses all input to the detector during the burst, the mean and variance of y(K) under the hypothesis H_1 for any permissible values c_1 and

q_1 are

$$E\{y(c_1, q_1, K) = Y_2 \leq Y_1 \tag{5}$$

and

$$\text{var}\{y(c_1, q_1, K)\} = V_2 \leq V_1 \tag{6}$$

It is obvious that in burst noise decisions based on the fixed level y_0 (y_0 is the optimum threshold in Gaussian noise) could result in significant errors of the second kind. Of course, threshold levels lower than y_0 would increase the error of the first kind in Gaussian noise. The fact that a priori knowledge of the burst width may be made available using simple threshold logic circuitry suggests the use of a variable threshold level detection scheme which in effect removes the dependence of Eqs. (5) and (6) on q and thus decreases the errors.

From practical considerations, the detection procedure could be implemented in terms of a number of fixed levels rather than a single variable level. Implementations of burst suppressor and multi-threshold detectors are shown in Figs. 1 and 2, respectively.

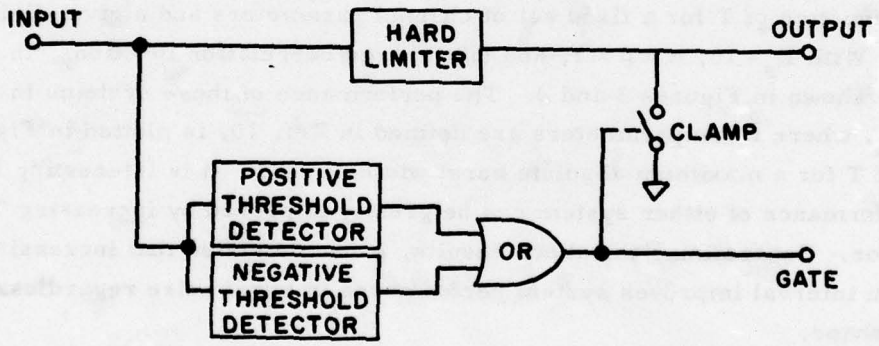


Fig. 1. Functional diagram of a burst suppressor.

B. Effect of Transmission Interval on System Performance

In considering the effect of a variable transmission interval on system performance in burst noise, two factors become immediately apparent. The length of the transmission interval $T = \Delta K$ and the channel bandwidth are directly related through the channel input-output state difference equation (see Ref. 10) so that for a given fixed input signal energy, the transmitted energy is directly proportional to T. In addition, for given maximum burst width, the effective width is inversely proportional to the length of the interval T.

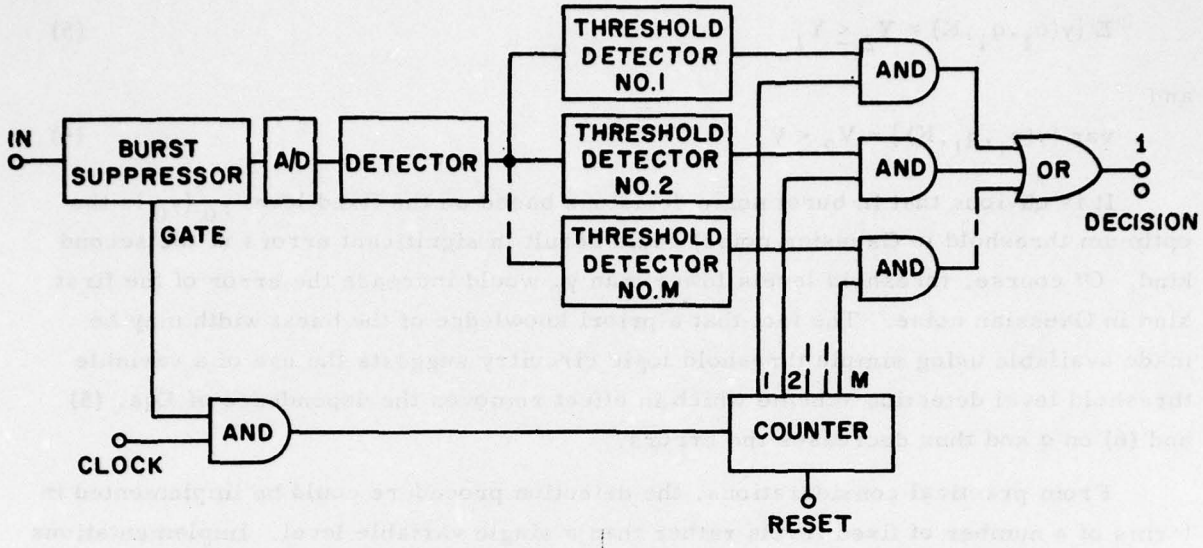


Fig. 2. Functional block diagram of a receiver containing a multi-threshold-level decision circuit.

As a numerical example, consider the performance of the systems of Refs. 10 and 11 as a function of T for a fixed set of channel parameters and a given finite signal energy. With $E_s = 10$, $\alpha = \beta = 1$, and unit energy correlation functions, the two systems are shown in Figures 3 and 4. The performance of these systems in terms of ρ_0 and ρ_{max}, where these parameters are defined in Ref. 10, is plotted in Fig. 5 as a function of T for a maximum absolute burst width of .433. It is interesting to note that the performance of either system can be greatly improved by increasing T even by a small factor. Generalizing the above results, it can be stated that increasing the transmission interval improves system performance in burst noise regardless of the signal waveshape.

Since a longer transmission interval corresponds to a lower data rate, a compromise between the value of T and the magnitude of the noise-to-signal ratio is required. It should be noted that for a given T, the data rate can be increased by employing other modes of transmission such as ternary, quaternary, or, in general, M-ary. In the case of M-ary transmission over a bandlimited channel, the optimal number of signals is related to the transmission interval, the noise-to-signal ratio, and the complexity of the receiver. This problem is discussed in detail elsewhere in this report. In the next section, an indication will be given about the influence of burst noise on the performance of M-ary system in gaussian noise and intersymbol interference.

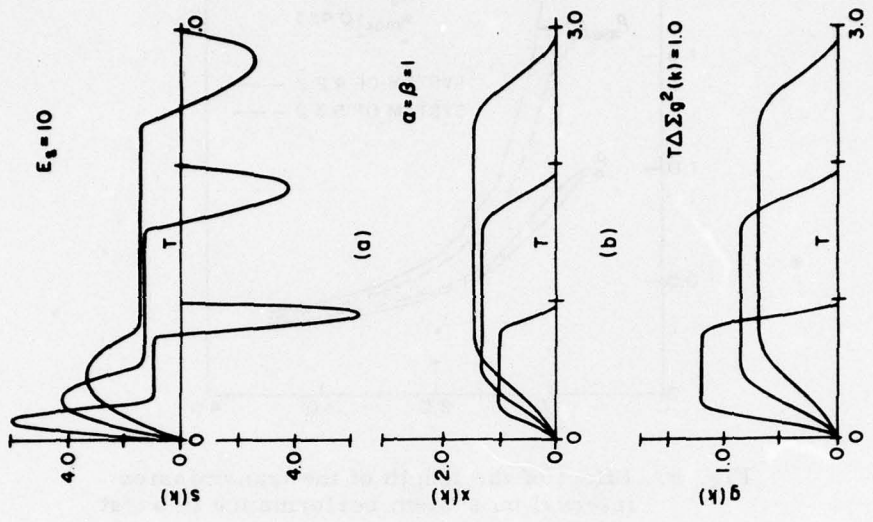


Fig. 3. Envelope of the discrete solutions as a function of the signal interval T.

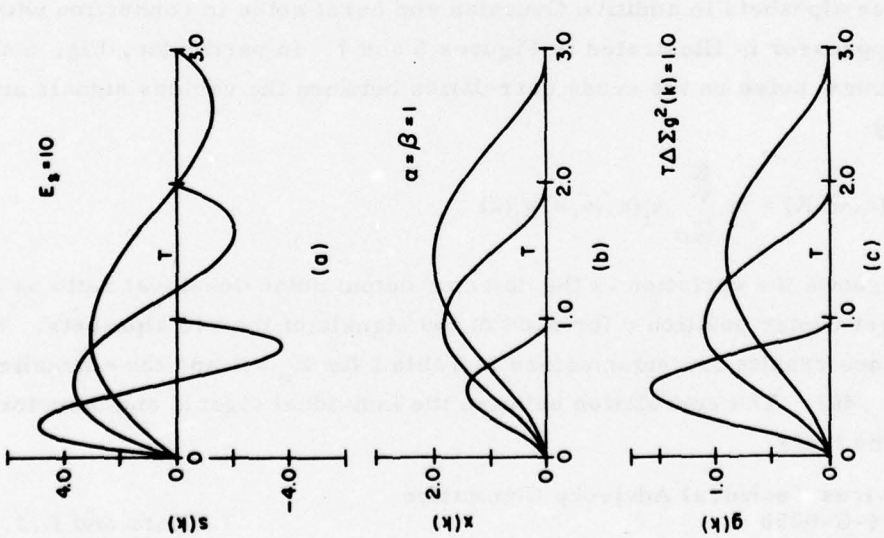


Fig. 4. Envelope of the discrete solutions as a function of the signal interval T.

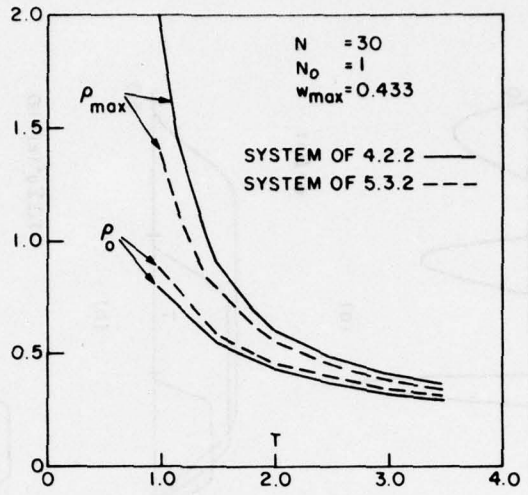


Fig. 5. Effect of the length of the transmission interval on system performance in burst noise.

C. The Influence of Burst Noise on the Performance of M-ary Systems in Gaussian Noise and Intersymbol Interference

Consider two quaternary alphabets based on Hadamard matrix transformation (see the presentation on M-ary detection in this report). The performance of systems using these alphabets in additive Gaussian and burst noise in connection with a non-linear suppressor is illustrated in Figures 6 and 7. In particular, Fig. 6 shows the effect of burst noise on the cross correlation between the various signals and detectors defined by

$$C_{ij}(c,w,K) = \Delta \sum_{k=o}^{K} x_i(c,w,k)\, g_j(k) \qquad (7)$$

Figure 7 shows the variation in the detector output noise-to-signal ratio as a function of the burst center position c for each of the signals of the two alphabets. The above performance results are summarized in Table I for $N_o = 1$ and the normalized burst width w = .467. The correlation between the individual signals and detectors is also given in the table.

Joint Services Technical Advisory Committee
F44620-74-C-0056
ERDA

L. Kurz and R.J. Nawrocky

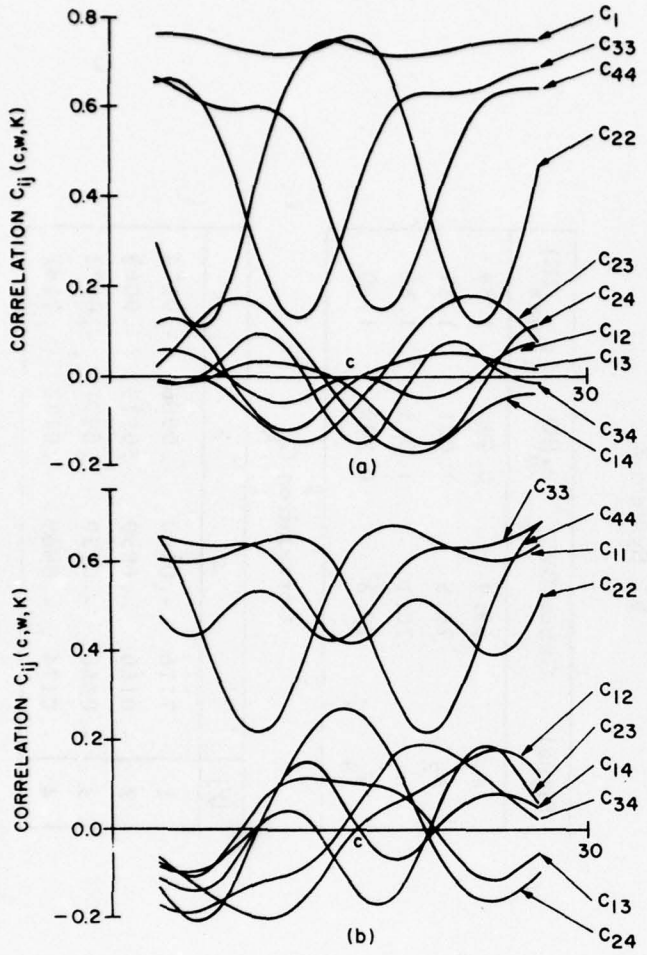


Fig. 6. Correlation (7) as a function of burst center position c for w = 0.467 and for the quaternary alphabets.

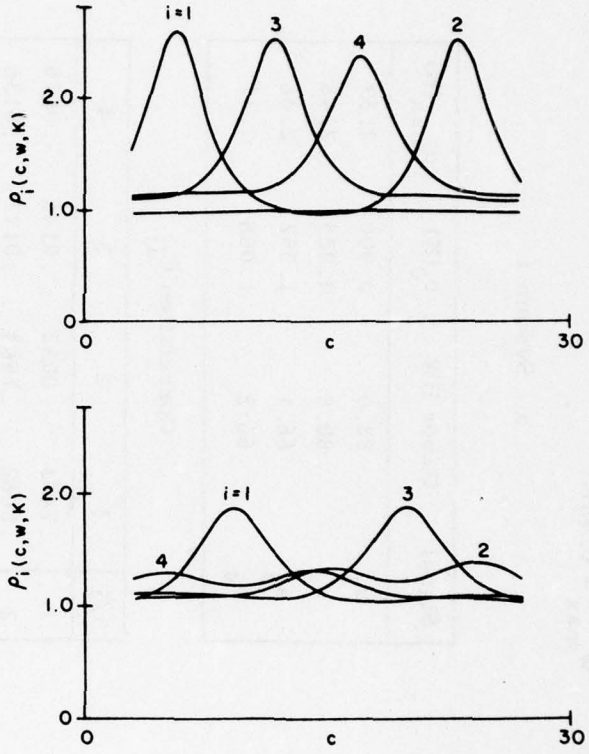


Fig. 7. Performance of transformed systems in additive Gaussian and burst noise for $N_0 = 1$, $w_{max} = 0.467$ and for the two quaternary alphabets.

TABLE I. Performance of Hadamard transformed quaternary orthogonal systems in additive Gaussian and burst noise; $B_s = 10$, $N_0 = 1$, $w_{max} = 0.467$.

a. System 1

Signal	Gabor BW	$\rho_0(K)$	$\rho_{max}(K)$
s_1	82.0	0.990	2.53
s_2	80.8	0.985	2.46
s_3	66.1	1.052	2.46
s_4	66.2	1.068	2.31

Correlation C_{ij}

i/j	1	2	3	4
1	.7453	.0062	.0146	-.0176
2	.0062	.7564	.0171	-.0136
3	.0146	.0171	.6746	-.0053
4	-.0176	-.0136	-.0053	.6589

b. System 2

Signal	Gabor BW	$\rho_0(K)$	$\rho_{max}(K)$
s_1	69.9	0.994	1.84
s_2	75.5	1.021	1.37
s_3	70.1	1.013	1.86
s_4	75.8	0.990	1.30

Correlation C_{ij}

i/j	1	2	3	4
1	.7116	-.0160	-.0066	-.0174
2	-.0160	.6950	-.0179	-.0069
3	-.0066	-.0179	.6977	-.0172
4	-.0174	-.0069	-.0172	.7141

REFERENCES

1. H.I. Silver and L. Kurz, "A Class of Continuous Signal Design Problems in Burst Noise," IEEE Trans. on Information Theory, IT-16, 570-581 (September 1970).
2. L. Kurz, "A Method of Digital Signalling in the Presence of Additive Gaussian and Impulsive Noise," 1962 IRE Int. Conv. Record, Pt. 4, 10, 161-173.
3. S.S. Rappaport and L. Kurz, "Optimal Decision Thresholds for Digital Signalling in Nongaussian Noise," 1965 IEEE Int. Conv. Record, Pt. 2, 13, 198-212.
4. S.S. Pappaport and L. Kurz, "An Optimal Nonlinear Detector for Digital Data Transmission Through Nongaussian Channels," IEEE Trans. on Communication Tech., COM-14, No. 3, 266-274 (June 1966).
5. J. Kapp and L. Kurz, "Performance of Two Suboptimum Detectors and Signal Selection in Gaussian and Impulsive Noise," Proc. 7th Annual Allerton Conference on Circuit and System Theory, 459-466 (October 1969).
6. D. Sheinbein and L. Kurz, "A Suboptimal Detector for a Class of Nongaussian Noise," 1972 International Symposium on Information Theory.
7. H.I. Silver and L. Kurz, "A Class of Discrete Signal Design Problems in Burst Noise," IEEE Trans. on Information Theory, IT-18, No. 2, 258-262 (March 1972).
8. D.C. Trimble and L. Kurz, "A Class of Jointly Optimal Signals and Detectors in Burst Noise," Proc. Sixth Annual Princeton Conference on Information Sciences and Systems, 252-256 (March 1972).
9. J.H. Miller and J.B. Thomas, "The Detection of Signals in Impulsive Noise Modeled as a Mixture Process," IEEE Trans. on Communications, COM-24, No. 5, 559-563 (May 1976).
10. L. Kurz and R.J. Nawrocky, "Data Transmission Link Design in Additive Gaussian and Burst Noise," *Progress Report No. 40* to JSTAC, Polytech. Inst. of New York, Report No. R-452.40-75, 410-418 (1975).
11. R.J. Nawrocky and L. Kurz, "Design of Digital Filters for Extraction of Signals from Intersymbol Interference and Noise," Proc. 8th Annual Princeton Conf. on Information Sciences and Systems, 499-503 (March 1974).

ON ESTIMATION OF CHANNEL CONDITIONS IN DIGITAL COMMUNICATION SYSTEMS

L. Kurz

The field of estimation of channel conditions or channel monitoring entails the estimation of the error rate in telemetry and data communication systems. The approaches fall into two categories which are designated as direct (estimating the actual error rate)[1,2] and extrapolated (estimation of a parameter of a known distribution).[3,4,5] The effort here is concentrated on a direct method useful in an M-ary (M > 2) data transmission or telemetry system. A nonparametric estimator of noise conditions, based on the output from M matched filters of which at a given time usually one is active (signal and noise) and M-1 are passive (just noise), is introduced. The proposed estimator is based on intuitive reasoning and is shown to have good qualities in estimating the true average probability of error. An expression for the mean of the estimator is developed in terms of the signal and noise distribution, and it is shown to be close to the true average probability of error.

A. Channel Description

For M-ary transmission the transmitter sends one of M-signals $(S_1, S_2, \cdots, S_M)$. The receiver is composed of M matched filters each of which is matched to one of the signals. At the end of each signaling interval, the outputs from each of the M filters are sampled. If the filter matched to S_i has the largest sampled output, the signal classifier assumes that S_i had been transmitted. If S_i is transmitted with a priori probability p_i, the sampled output of the k-th filter $(k \neq i)$ is equal to n_k (just noise), while the sampled output of the i-th filter is equal to $S_i + n_i$. It is assumed that the p.d.f.'s of all the idle and active filters are independent, the p.d.f.'s of all the idle filters are the same, and the p.d.f. of the active filter is the same independently of which signal was sent. Hence,

$$f_{n_j, n_k}(\xi, \eta) = f_{n_j}(\xi)\, f_{n_k}(\eta) \qquad (j \neq k)$$

where

$f_{n_j}(\cdot)$ = p.d.f. of the j-th idle filter

$f_{n_k}(\cdot)$ = p.d.f. of the k-th idle filter

and

$f_{n_j, n_k}(\cdot, \cdot)$ = joint p.d.f. of the j-th and k-th idle filters

$$f_{n_j}(\cdot) = f_{n_k}(\cdot) = f_n(\cdot)$$

and

$$f_{sig_i}(x) = f_{sig_j}(x) \quad \text{for all i and j}$$

where

$$f_{sig_i}(x) = f_{sig}(x) = \text{p.d.f. of } S_i + n_i, \text{ for all i.}$$

Hereafter attention will be given to the combined output $(S_i + n_i)$ rather than individual outputs S_i and n_i. The only necessary assumption is that the output of $S_i + n_i$ be independent of the output n_k $(k \neq i)$. Therefore, the assumption that the noise output in the active filter be independent of the signal output in the active filter is unnecessary, and the solution to the monitoring problem may be extended to nonlinear processing by the receiver.

B. Expression for the Probability of Error

The probability of error is equal to the probability that one of M - 1 independent noise samples will be greater than the output of signal plus noise. (One should note that since the probability of error is the same, independently of which signal was sent, the probability of error will be independent of p_i.) The probability of error is equal to the probability that the signal will be equal to x, multiplied by the probability that the maximum of the noise samples is greater than x, summed over all possible values of x or

$$P_e = \int_{-\infty}^{\infty} f_{sig}(x) \int_{x}^{\infty} f_{max}(y)\, dy\, dx \tag{1}$$

where

$f_{max}(y)$ = p.d.f. of the maximum of M - 1 independent noise samples

It should be noted that for M = 2, Eq. (1) reduces to

$$P_e = \int_{-\infty}^{\infty} f_{sig}(x) \left[\int_{x}^{\infty} f_n(y)\, dy\right] dx$$

$$= 1 - \int_{-\infty}^{\infty} f_{sig}(x)\, F_n(x)\, dx \tag{2}$$

Since the probability of error is small, the maximum value of the sampled outputs of the idle filter nearly represent the maximum value of the M - 1 independent noise samples, and the sampled outputs of the active filter nearly represent samples of signal plus noise. The samples are then taken from the maximum of the idle filters

and from the active filter, and a linear statistic is calculated using the values of the two outputs.

The statistic used in estimating the average probability of error is a modified version of the Mann-Whitney test.[6] Let $x_1, x_2, \cdots, x_k$ represent k sampled outputs of the maximum of idle filters, and $y_1, y_2, \cdots, y_k$ represent the k corresponding outputs from the active filter. The statistic used is

$$P_{e_m} = \frac{W' + k(k-1)}{2k(k-1)} \tag{3}$$

where

$$W' = \sum_{\substack{i=1 \\ i \neq j}}^{k} \sum_{j=1}^{k} \operatorname{sgn}(x_i - y_j) \tag{4}$$

and the standard Mann-Whitney test is

$$W = \frac{1}{k^2} W'$$

The choice of the estimator given by Eq. (3) is not arbitrary but follows from logical reasoning. For the ordinary Mann-Whitney test, W, it is well known that the mean is

$$E[W] = P^+ - P^-$$

where

$$P^+ = \text{Prob}[x > y] \quad \text{and} \quad P^- = \text{Prob}[x < y] \tag{5}$$

Since x and y have continuous c.d.f.'s, $P(x = y) = 0$ and $P^+ + P^- = 1$. P^+ can be expressed in terms of E[W] as

$$E[W] = [P^+ - (1 - P^+)] = [2P^+ - 1] \tag{6}$$

and

$$P^+ = \frac{1 + E[W]}{2} \tag{7}$$

Since P^+ is the quantity which is sought, the measured value of W is substituted for its mean to obtain an estimator of the average probability of error.

Since the estimate of P_{e_m} is in the form $aW' + b$ and W' is asymptotically normal, it follows that P_{e_m} is asymptotically normal. The random variable P_{e_m} is, therefore, completely specified by its mean and variance, with the mean satisfying

$$E[P_{e_m}] = P^{+}_{idle} \tag{8}$$

where

$$P^{+}_{idle} = \text{Prob}[x_i > y_i] \quad i \neq j \tag{9}$$

Since x_i is a sample of the idle filter and y_j is a sample of the active filter, it is necessary to find the p.d.f. of each of the above two filters. Denoting the p.d.f.'s of the two filters by $f_{idle}(x)$ and $f_{act}(x)$, respectively, one obtains

$f_{idle}(x)$ = p.d.f. of min {a sample of f_{max}, a sample of $f_{sig}(x)$}

and

$f_{act}(x)$ = p.d.f. of max {a sample of f_{max}, a sample of $f_{sig}(x)$}

It is eash to show that

$$F_{max}(x) = F_n^{m-1}(x) \tag{10}$$

$$F_{act}(x) = F_{sig}(x)\, F_n^{m-1}(x)$$

$$f_{act}(x) = [F_{sig}(x)]\,[m-1]\, F_n^{m-2}(x)\, f_n(x) + f_{sig}(x)\, F_n^{m-1}(x) \tag{11}$$

Similarly

$$F_{idle}(x) = F_{sig}(x) + F_n^{m-1}(x) - F_{sig}(x)\, F_n^{m-1}(x)\, F_n^{m-1}(x) \tag{12}$$

and

$$f_{idle}(x) = f_{sig}(x) + (m-1)\, F_n^{m-2}(x)\, f_n(x)$$

$$- (m-1)\, F_{sig}(x)\, F_n^{m-2}(x)\, f_n(x) + f_{sig}(x)\, F_n^{m-1}(x) \tag{13}$$

Since

$$P^{+}_{idle} = \int_{-\infty}^{\infty} f_{act}(x)\,[1 - F_{idle}(x)]\,dx = 1 - \int_{-\infty}^{\infty} f_{act}(x)\, F_{idle}(x)\,dx$$

combining Eqs. (12) and (13), yields

$$P^{+}_{idle} = 1 - \int_{-\infty}^{\infty} [F_{sig}(x) + F_n^{m-1}(x) - F_{sig}(x)\, F_n^{m-1}(x)]\,[F_{sig}(x)(m-1)\, F_n^{m-2}(x)\, f_n(x)$$

$$- f_{sig}(x)\, F_n^{m-1}(x)]\,dx$$

Figures 1-3 represent typical performance curves for indicated signal and noise conditions. From the curves it follows that the estimator suggested here compares

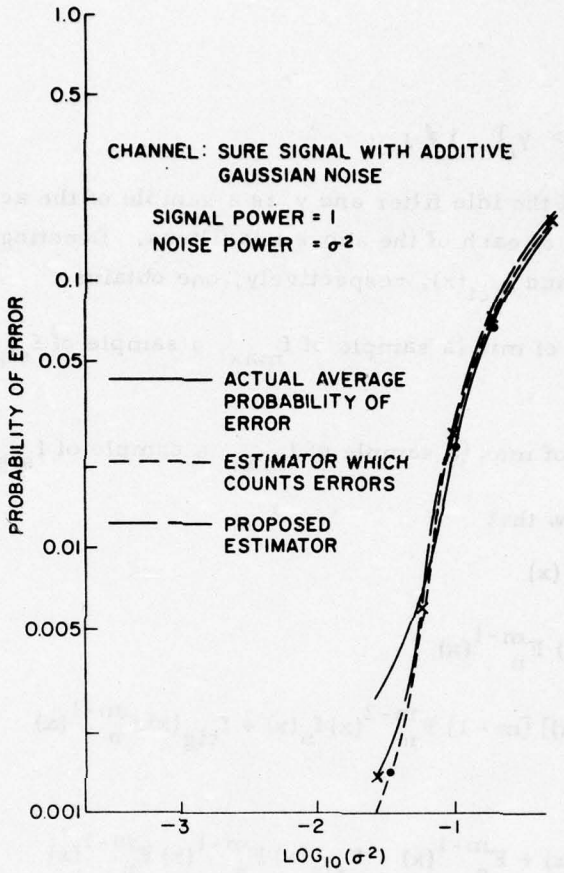


Fig. 1. Estimators of probability of error as a *function of equivalent signal-noise ratio* parameter.

favorably with the theoretical and counted error rate for three different classes of channels.

Joint Services Technical Advisory Committee
F44620-74-C-0056

L. Kurz

REFERENCES

1. M.I. Epstein and L. Kurz, "Estimators of Channel Conditions in the Presence of Interchannel Interference," Proc. IEEE Intl. Conf. on Communications, 46.1-46.5 (June 1972).

2. J.C. Ashlock and E.C. Posner, "Application of Statistical Theory of Extreme Values to Space Craft Receivers," JPL Lab., Pasadena, Ca., Technical Report 32-737 (May 1965).

3. D. Chesler and G. Hingorami, "Performance of Probability Measuring Unit for Flat Fading Channel," Sylvania Applied Research Lab., RN 545 (March 1965).

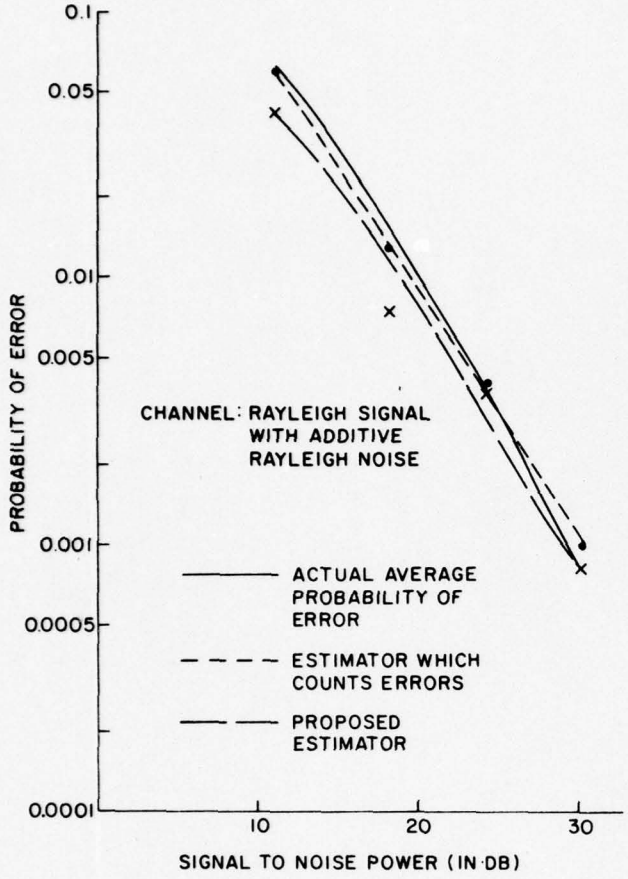


Fig. 2. Estimators of probability of error as a function of equivalent signal-noise ratio parameter.

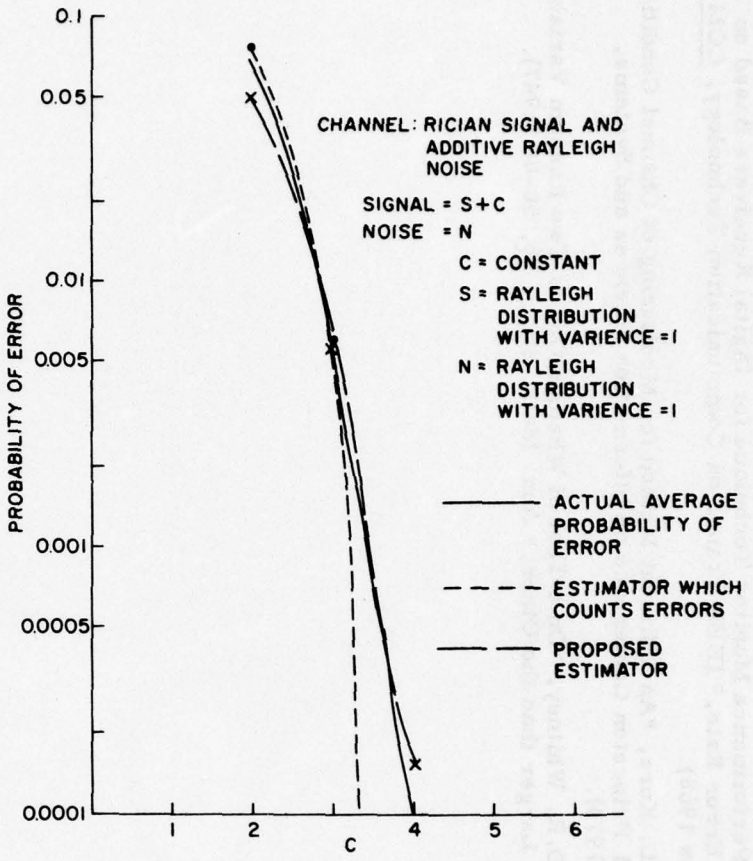


Fig. 3. Estimators of probability of error of a function of equivalent signal-noise ratio parameter.

4. D.J. Gooding, "Performance Monitor Technique for Digital Receivers Based on Extrapolation of Error Rate," IEEE Trans. on Communication Technology, COM-16, 380-387 (June 1968).

5. J.G. Evans and L. Kurz, "An Efficient Method for Monitoring of Channel Conditions," Proc. 8th Annual Princeton Conference on Information Sciences and Systems, 510-513 (March 1974).

6. H.B. Mann and D.R. Whitney, "On a Test of Whether One of Two Random Variables is Stochastically Larger than the Other," Ann. Math. Stat., 18, 50-60 (1947).

A ROBUST VECTOR ROBBINS-MONRO PROCEDURE WITH APPLICATION TO PARAMETER ESTIMATION IN m-INTERVAL DETECTORS

P. Kersten and L. Kurz

In a recent report[7] the use of the Robbins-Monro (R-M) and the Kiefer-Wolfowitz (K-W) procedure to create robust recursive estimators was illustrated. In both of these scalar estimation procedures, the rate of convergence is dependent upon the choice of the gain coefficient which in part controls the step size of this iterative corrective procedure. To enhance the rate of convergence, adaptive gain coefficients which closely estimate the optimum gain coefficients are inserted. In addition, by using non-linear regression functions, these recursive estimators can be made robust. Both of these concepts may be extended to produce similar results in the vector R-M procedure.

Like the scalar version of the R-M procedure, several authors have considered the convergence of the vector R-M procedure. Dvoretzky,[5] Derman[4] and Sacks[10,4] have considered one or more forms of convergence of the vector R-M procedure. Under suitable conditions the vector R-M procedure converges in mean-square, wpl, and in law. In Section A, a theorem which establishes the asymptotic normality of the vector R-M procedure with adaptive gain matrices is stated. This is the vector analog of Theorem 1 of Reference 7. In Section B, an application of this procedure to parameter estimation in m-interval detectors is presented. It is shown how adaptive gain matrices are used in the estimation of the quartiles of a distribution. The resulting recursive estimator is not only computationally attractive, but also robust.

A. The Vector Robbins-Monro Procedure

The usual version of the vector Robbins-Monro procedure (hereafter referred to as the R-M procedure) is given by the recursion equation

$$\underline{X}_{n+1} = \underline{X}_n - a_n(\underline{Y}(\underline{X}_n) - \underline{\alpha})$$

where $\underline{X}_n$, $\underline{X}_{n+1}$, $\underline{Y}(\underline{X}_n)$ and $\underline{\alpha}$ are q-dimensional vectors and $\{a_n\}$ is a sequence of positive scalar constants s.t. $\sum a_n = \infty$ and $\sum a_n^2 < \infty$. If the stochastic approximation (S.A.) procedure is visualized as q distinct (possibly dependent) scalar S.A. processes instead of a single q-dimensional process, the scalar gain coefficient should be replaced by a gain matrix. The advantage of this is to "match" the diagonal entries in the gain matrix to the q distinct S.A. procedures. If the gain coefficient a_n is adaptive, then a_n generalizes to a random gain matrix which we denote as $A_n(\cdot) = A_n(\underline{X}_1 \cdots, \underline{X}_n)$. Accordingly, the conditions $\sum a_n = \infty$ and $\sum a_n^2 < \infty$ must now be replaced with similar conditions holding wpl upon the eigenvalues of the gain matrix $A_n(\cdot)$. We adopt the Euclidean norm as our metric and the corresponding matrix norm that it induces, i.e., the spectral norm.[9]

The R-M procedure is defined by the recursion equation

$$\underline{X}_{n+1} = \underline{X}_n - \frac{1}{n} A_n(\underline{X}_1, \cdots, \underline{X}_n)\,[\underline{Y}(\underline{X}_n; \underline{\theta}) - \underline{a}]$$

where $\underline{X}_n$, $\underline{Y}(\underline{X}_n; \underline{\theta})$, $\underline{\theta}$, and $\underline{a}$ are all q-dimensional vectors, $E\underline{Y}(\underline{X}; \theta) = \underline{M}(\underline{X})$ and $\underline{M}(\underline{X}) = \underline{a}$ has the unique solution $\underline{X} = \underline{\theta}$. The adaptive gain matrices A_n are assumed to be positive definitive with eigenvalues $a_1^{(n)}(\cdot) > \cdots > a_q^{(n)}(\cdot) > 0$ s.t. $\sum a_q^{(n)}(\cdot)/n = \infty$ wpl. The assumptions follow closely those of Sacks[10] in spirit and form.

The first three assumptions establish the relationship of the i.i.d. random vectors $\underline{Z}_n(\underline{X})$ to the regression function, namely, the noise is additive and zero mean.

(1) $E[\underline{Y}(\underline{X}_n; \underline{\theta}) | \underline{X}_n, \cdots, \underline{X}_1] = E[\underline{Y}(\underline{X}_n; \underline{\theta}) | \underline{X}_n] = \underline{M}(\underline{X}_n; \underline{\theta})$

abbreviated $\underline{M}_n$.

(2) $\underline{Y}(\underline{X}_n; \underline{\theta}) = \underline{M}(\underline{X}_n; \underline{\theta}) + \underline{Z}(\underline{X}_n) = \underline{M}(\underline{X}_n) + \underline{Z}(\underline{X}_n) = \underline{M}_n + \underline{Z}_n$ where

$\underline{E}\,\underline{Z}_n = E\underline{Z}_n(\underline{X}_n) = E(\underline{Z}_n | \underline{X}_n) = E(\underline{Z}_n | \underline{X}_n, \cdots, \underline{X}_1) = 0$ wpl.

(3) $\underline{M}(\underline{X}_n; \underline{\theta}) = \underline{a}$ has a unique root i.e., $\underline{X} = \underline{\theta}$.

To show convergence, the class of regression functions must be restricted in the same manner as the scalar case. The following three assumptions serve this purpose.

(4) $(\underline{X} - \underline{\theta})^T (\underline{M}(\underline{X}, \underline{\theta}) - \underline{a}) > 0$ for all $\underline{X} \neq \underline{\theta}$.

(5) $K\|\underline{X} - \underline{\theta}\| \leq \|\underline{M}(\underline{X}; \theta) - \underline{a}\| \leq K_1 \|\underline{X} - \underline{\theta}\|$ for $\|\underline{X} - \underline{0}\| < \eta$ and

$K\eta \leq \|M(\underline{X}, \underline{\theta}) - \underline{a}\| \leq K_1 \|\underline{X} - \underline{\theta}\|$ for $\eta < \|\underline{X} - \underline{\theta}\| < \infty$

(6) $\underline{M}(\underline{X}, \underline{\theta}) = \underline{a} + B(\underline{X} - \underline{\theta}) + \underline{\delta}(\underline{X}, \underline{\theta})$ where B is a positive definite q x q matrix s.t. $\|B\| < \infty$ and $\|\underline{\delta}(\underline{X}_n, \underline{\theta})\| = 0(\|\underline{X}_n - \underline{\theta}\|)$ as $\underline{X} - \underline{\theta} \to \underline{0}$.

Further, there exists an orthogonal transformation P. s.t. $PP^T = I$, $P^T B P$ = diagonal matrix and $P^T A_n P$ = diagonal matrix for each n.

As in the scalar case, the additive noise vector $\underline{Z}(\underline{X})$ must have a uniformly bounded variance and have a well defined covariance matrix as $\underline{X} \to \underline{\theta}$ a.s. In addition, the adaptive gain matrix needs to be well behaved and a consistent "mean square" estimator of a constant matrix. These requirements are reflected in the following assumptions:

(7) (a) $\sup_{\underline{X}} E\|\underline{Z}(\underline{X})\|^{2+2\varepsilon} < \infty$ for some $\varepsilon > 0$

(b) $\lim_{\underline{X} \to \underline{\theta}} E[\underline{Z}(\underline{X})\underline{Z}^T(\underline{X})] = \pi$ where π is a non-negative definite matrix and

where the limit is in the sense of the norm.

(8) $a_1^{(n)} \geq a_2^{(n)} \geq \cdots \geq a_q^{(n)} > 0$ be the eigenvalues of A_n and $b_1 \geq b_2 \geq \cdots \geq b_q > 0$ be the eigenvalues of B and $a_1 \geq a_2 \geq \cdots \geq a_q > 0$ be the eigenvalues of A

(a) $0 < a_1' \leq \inf_{\underline{X}} \|A_n\| \leq \sup_{\underline{X}} \|A_n\| \leq a'' < \infty$ wpl for n large

(b) $\lim_n a_q^{(n)}(\cdot) \geq a_1' > 0$ wpl where A is a constant matrix s.t. $a' \leq \|A\| \leq a''$.

(9) $a_1' b_q - \varepsilon > 1/2$ and $a_q b_q - \varepsilon > 1/2$

Assumption (9) has no readily apparent physical interpretation other than it is required for convergence within the proof.

Theorem 1: Under assumptions (1) - (9), let $r_1 > r_2 > \cdots > r_q > 0$ be the eigenvalues of AB. Then $n^{1/2}(\underline{X}_n - \underline{\theta})$ is asymptotically normal with mean zero and covariance matrix PQP^T where Q is a matrix whose $(i,j)^{th}$ element is $a_i \pi^*_{ij} a_j [r_i + r_j - 1]^{-1}$ where $\pi^* = P^{-1} \pi P$.

Proof: See appendix of Reference 1.

B. The Simultaneous Estimation of the Quartiles of a Distribution Function

In this example, the R-M procedure is used for the simultaneous estimation of the quartiles of a symmetric distribution given random samples from the distribution. It will become clear that this procedure generalizes to the simultaneous estimation of n-arbitrary quantiles. The use of adaptive gain matrices in conjunction with the R-M procedure ensures that this recursive estimator converges at a "near" optimum rate. For real time estimation of quantiles this becomes an important consideration. For instance, the nonparametric m-interval partition detector considered by Ching and Kurz[3] requires the estimation of a finite number of quantiles of the noise distribution for its implementation. To enable this particularly attractive detector to adapt to a slowly-varying noise distribution, training samples can be taken periodically to update the estimates of these quantiles. Since the R-M procedure is recursive, it requires little sample storage and is amenable to digital computation. In fact, it is ideally suited for this task. Clearly, the adaptiveness of the detector will be dependent in part upon the rate of convergence of the quantile estimation procedure. In communication channels where the noise distribution is a linear combination of a low variance noise distribution and a Burst Noise component, it is also necessary that this estimation procedure be robust. In this example, the robustness will be guaranteed by the fact that our regression function is bounded and non-linear, and our adaptive gain

matrices are based upon linear order statistics. It is this very problem which motivates the generality of Theorem 1.

Before proceeding to the three-dimensional case, the one-dimensional case will be examined to provide an easier setting in which to discuss the properties of the linear order statistics used in the adaptive gain matrices. In the scalar case, the recursive equation becomes

$$X_{p+1} = X_p - \frac{1}{p} A_p S_m(X_p, \lambda_k)$$

where

$$S_m(X_p, \lambda_k) = \frac{1}{m} \sum_{i=1}^{m} \operatorname{sgn}[X_p - Y_{i+m(p-1)}] + (1 - 2\lambda_k)$$

and

$$A_p = \frac{m+1}{2(p-1)} \sum_{j=1}^{p-1} [Z_{j,[m\lambda_k]+1} - Z_{j,[m\lambda_k]}]$$

where

$$Z_{j,[m\lambda_k]+1} \triangleq [m\lambda_k] + 1^{th}$$

order statistics from the random samples $Y_{i+m(p-i)}$, $i = 1, \cdots, m$ with CDF $F(x)$ and $[m\lambda_k]$ defined to be the greatest integer less than or equal to $m\lambda_k$.

The estimator A_p is essentially the same one used in Ref. 6 and is a biased estimator of the optimum gain coefficient $1/2f(\theta_k)$ where θ_k is the λ_k^{th} quantile. Since A_p is an average of i.i.d. random variables, if the $E|Z_{j,[m\lambda_k]+1} - Z_{j,[m\lambda_k]}| < \infty$ then A_p converges a.s. to $E[Z_{j,[m\lambda_k]+1} - Z_{j,[m\lambda_k]}]$ via the Strong Law of Large Numbers. Thus to study the limit of A_p is to evaluate the expected value of A_p, assuming it is finite. Unfortunately, the latter task is in general difficult and one must resort to numerical methods of integration to obtain accurate answers. More importantly, in the one-dimensional case, one must be careful that $A_p \not\to 1/(4f(\theta_k)) = A$ a.s. since it is well known that $\sqrt{p}\, X_p \sim N(\theta_k, A^2\sigma^2/(2A\alpha - 1))$ where $\alpha = 2f(\theta_k)$ and $\sigma^2 = \text{Var } S_m$. If $A_p \to 1/(4f(\theta_k))$ a.s. then the variance is infinite and the convergence is meaningless. This variance will be minimum if $A\alpha = 1$ (Reference 8). If A_p is biased, it is best that it be greater than $1/(4f(\theta_k))$ so that the rate of convergence is "close" to optimal. To do this a lower bound is established on the estimator $\frac{n+1}{2}[Z_{(i+1)} - Z_{(i)}]$.

For convenience, consider $\delta_i = (n+1)[Z_{(i+1)} - Z_{(i)}]$, the estimator of $1/f(\theta)$. Its expected value is given by

$$E\delta_i = (n+1)\, E[Z_{(i+1)} - Z_{(i)}] = (n+1) \int_0^1 G(x)\, [\beta_{i+1}(x,n) - \beta_i(x,n)]\, dx$$

where

$$\beta_i(x,n) = \frac{n!}{(i-1)!\,(n-1)!}\, x^i (1-x)^{n-i}$$

the PDF of the ith order statistic from a uniform random sample of size n and

$$G(x) = \inf_y \{y : F(y) \geq x\} = F^{-1}(x)$$

F(x) being the CDF of the random samples. We now defined a class of distribution functions $\mathcal{J}(i)$ as those CDF's which possess the following two properties.

(1) $\lim F^{-1}(x)\, \beta_{i+1}(x, n+1) = 0$ as $x \to 0+$ and $x \to 1-$ and

(2) $G'(x) = 1/f(F^{-1}(x))$ is a convex function for $x \in (0, 1)$.

The first property allows us to evaluate the expectation of δ_i via integration by parts as

$$E\delta_i = -F^{-1}(x)\, \beta_{i+1}(x,n+1)\Big|_0^1 + \int_0^1 G'(x)\, \beta_{i+1}(x,n+1)\, dx = \int_0^1 \frac{1}{f(F^{-1}(x))}\, \beta_{i+1}(x,n+1)\, dx$$

The second property permits us to apply Jensen's Inequality to conclude that

$$E\delta_i = EG'(X) \geq G'(EX) = G'\left(\frac{i+1}{n+2}\right) = 1/f\left(F^{-1}\left(\frac{i+1}{n+2}\right)\right)$$

The resulting class of functions $\mathcal{J}(i)$ contains the Gaussian, Cauchy and Laplace distributions and this lower bound ensures that $(1/2)E\delta_i$ is strictly greater than $1/(4f(\theta_i))$ for these distributions. Reference to Table I will confirm that this bound is conservative and the calculated values of $\frac{n+1}{2} E\delta_i = \frac{n+1}{2} E[Z_{j,i+1} - Z_{j,i}]$ all exceed $1/2\, f(\theta_i)$ for the three aforementioned distributions. However, if the distribution is unknown except for the fact that $f(x) \leq f(y)$ for $|x| > |y|$ and $F(x) \in \mathcal{J}(i-1)$, then one can conclude that the estimator $Z_{j,i} - Z_{j,i-1}$ instead of $Z_{j,i+1} - Z_{j,i}$ will have a lower bound of $1/2\, f(F^{-1}(\frac{i}{n+2})) \geq 1/2\, f(F^{-1}(i/n)) = 1/2\, f(\theta_i)$ for $\theta_i \leq F^{-1}(1/2)$, which suffices for symmetric density functions. This provides a safer estimate but in view of Table I is unwarranted for this particular example.

TABLE I

Quartile	First Quartile i = 4			Second Quartile i = 8			Third Quartile i = 12		
Distribution	$E\delta_i/2$	$1/2\,f(\theta_i)$	Lower Bound	$E\delta_i/2$	$1/2\,f(\theta_i)$	Lower Bound	$E\delta_i/2$	$1/2\,f(\theta_i)$	Lower Bound
Laplace	2.129732	2.0	1.8	1.270854	1.0	1.0	2.129723	2.0	1.8
Gaussian	1.642492	1.573113	1.49	1.312659	1.255314	1.253314	1.642452	1.573113	1.49
Cauchy	4.247329	3.141593	2.68	1.846653	1.570796	1.570796	4.247329	3.141593	2.68

Having established the properties of A_p for symmetric distributions, the extension to the 3-Dimensional case is straightforward. Toward this end, define the gain matrix by,

$$A_p = \begin{bmatrix} \frac{m+1}{2(p-1)} \sum_{j=1}^{p-1} (Z_{j,[\frac{m}{4}]+1} - Z_{j,[\frac{m}{4}]}) & 0 & 0 \\ 0 & \frac{m+1}{2(p-1)} \sum_{j=1}^{p-1} (Z_{j,[\frac{m}{2}]+1} - Z_{j,[\frac{m}{2}]}) & 0 \\ 0 & 0 & \frac{m+1}{2(p-1)} \sum_{j=1}^{p-1} (Z_{j,[\frac{3m}{4}]+1} - Z_{j,[\frac{3m}{4}]}) \end{bmatrix}$$

and the vector version of S_m by

$$S_M(\underline{X}_p, \underline{\lambda}) = \begin{bmatrix} \frac{1}{m} \sum_{i=1}^{m} \operatorname{sgn}[X_{p_1} - Y_{i+m(p-1)}] + (1 - 2\lambda_1) \\ \frac{1}{m} \sum_{i=1}^{m} \operatorname{sgn}[X_{p_2} - Y_{i+m(p-1)}] + (1 - 2\lambda_2) \\ \frac{1}{m} \sum_{i=1}^{m} \operatorname{sgn}[X_{p_3} - Y_{i+m(p-1)}] + (1 - 2\lambda_3) \end{bmatrix}$$

The regression function $\underline{M}(\underline{X}_p, \theta) = ES_m(\underline{X}_p, \underline{\lambda})$ is now given by

$$\underline{M}(\underline{X}_p, \underline{\theta}) = 2\begin{bmatrix} F(X_{p_1}) - \lambda_1 \\ F(X_{p_2}) - \lambda_2 \\ F(X_{p_3}) - \lambda_3 \end{bmatrix} = 2\begin{bmatrix} f(\theta_1) & 0 & 0 \\ 0 & f(\theta_2) & 0 \\ 0 & 0 & f(\theta_3) \end{bmatrix}\begin{bmatrix} X_{p_1} - \theta_1 \\ X_{p_2} - \theta_2 \\ X_{p_3} - \theta_3 \end{bmatrix}$$

$$+ \begin{bmatrix} f'(\theta_1) & & \\ 0 & f'(\theta_2) & \\ 0 & & f'(\theta_3) \end{bmatrix}\begin{bmatrix} (X_{p_1} - \theta_1)^2 \\ (X_{p_2} - \theta_2)^2 \\ (X_{p_3} - \theta_3)^2 \end{bmatrix} + \text{H.O.T.'s}$$

The B matrix is given by

$$B = \begin{bmatrix} f(\theta_1) & 0 & 0 \\ 0 & f(\theta_2) & 0 \\ 0 & 0 & f(\theta_3) \end{bmatrix}$$

and the additive noise vector $\underline{Z}(\underline{X}_p, \underline{\theta})$ is given by

$$\underline{Z}_p = \begin{bmatrix} \frac{1}{m}\sum_{i=1}^{m} \operatorname{sgn}(X_{p_1} - Y_{i+m(p-1)} + 1 - 2F(X_{p_1}) \\ \frac{1}{m}\sum_{i=1}^{m} \operatorname{sgn}(X_{p_2} - Y_{i+m(p-1)} + 1 - 2F(X_{p_2}) \\ \frac{1}{m}\sum_{i=1}^{m} \operatorname{sgn}(X_{p_3} - Y_{i+m(p-1)} + 1 - 2F(X_{p_3}) \end{bmatrix}$$

And by the result established, we are sure that $A_{ij} > \frac{\delta_{ij}}{4f(\theta_i)}$ and thus $[AB]_{ij} = r_i \delta_{ij}$ where $r_i > 1/2$.

With the above notation, conditions (1) and (2) are transparent and condition (3) is satisfied if F(x) is strictly increasing and continuous which we simply assume. If $\lambda_i = i/4, i = 1, 2, 3$ with corresponding quartiles θ_i, then

$$(\underline{X} - \underline{\theta})^T (\underline{M}(\underline{X}, \underline{\theta}) - \underline{a}) = \sum_{i=1}^{3} (X_i - \theta_i)(F(X_i) - \lambda_i) > 0 \quad \text{if} \quad X_i \neq \theta_i$$

since $X_i > \theta_i$ implies $F(X_i) > \lambda_i$ and $X_i < \theta_i$ implies $F(X_i) < \lambda_i$ by our assumption that

$F(x)$ be strictly increasing. To satisfy conditions (5) and (6), we assume $f(x)$ is continuous a.e. and its support is convex. Moreover, it is assumed that for all x, $f(x)$ is bounded $f(\theta_i) > 0$, and $0 < |f'(\theta_i)| < \infty$, $i = 1, 2, 3$. The last assumption along with the expansion of $\underline{M}(\underline{X}_p, \underline{\theta})$ guarantee condition (6). Condition (5) is also satisfied since

$$\|\underline{M}(\underline{x}) - \underline{a}\|^2 = \sum_{i=1}^{3} (F(x_i) - \lambda_i)^2$$

and the assumptions on $F(x)$ mean geometrically that each $F(X_i) - \lambda_i$ is bounded above and below as in Figure 1. Since A_p is diagonal, it follows that the matrix P specified after assumption (6) is taken as the identity matrix.

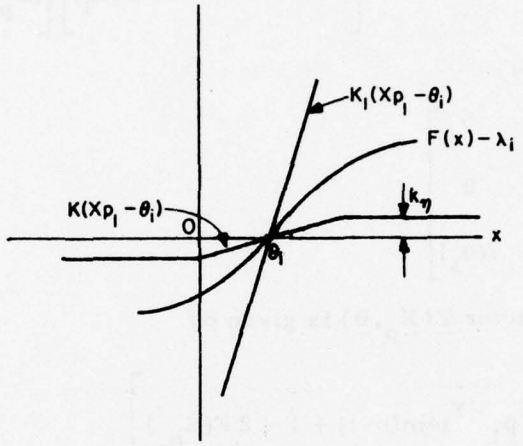


Fig. 1.

The sgn function which appears in each component of $\underline{Z}_p$ assures that $E\|\underline{Z}(x)\|^{2+2\varepsilon}$ is bounded uniformly in x which in turn implies condition (7a). The covariance matrix as $\underline{X} \to \underline{\theta}$ is calculated to be

$$[\pi]_{ij} = \frac{4}{m} \min(\lambda_i, \lambda_j)[1 - \max(\lambda_i, \lambda_j)]$$

and π is positive definite thus verifying condition (7b). The form of $[A_p]_{ii}$ allows one to state that $Z_{j',i+1} - Z_{j',i} > 0$ wpl by the fact that $F(x)$ is strictly increasing and continuous and therefore

$$\lim_{p} [A_p]_{ii} > 0 \text{ wpl for } 1, 2, 3$$

which yields condition (8b). The above argument establishes the first half of condition (8a) and the fact that $E[A_p]_{ii}$ is finite for all i suffices to establish that $\sup_x \|A_p\| < \infty$ wpl. Conditions for the existence of both the expectation and the variance for the order statistics is given in Ref. 2, page 44. For our particular example, we shall assume

$n = 16$, $i = 4$ and thus even if $F(x)$ is Cauchy, the expectation and variance of A_p are finite. Condition (8) requires the existence of a diagonal matrix A s.t. $\lim_{n\to\infty} E\|A_n - A\|^2 = 0$ or in other words

$$\lim_{p\to\infty} \max_i \left[\frac{m+1}{2(p-1)} \sum_{j=1}^{p-1} \left[Z_{j,[\frac{im}{4}]+1} - Z_{j,[\frac{im}{4}]}\right] - A_{ii}\right]^2 = 0,\ i = 1,2,3$$

Since $A_{ii} = E[A_p]_{ii}$ it suffices to show $\mathrm{Var}[A_p]_{ii} \to 0$ as $p \to \infty$ for all i. But

$$\mathrm{Var}[A_p]_{ii} = \frac{(m+1)^2}{4(p-1)} \mathrm{Var}\left[Z_{j,[\frac{im}{4}]+1} - Z_{j,[\frac{im}{4}]}\right] = 0\left(\frac{1}{p}\right)$$

provided the latter variance is bounded. Extension of this procedure to n arbitrary quantiles requires that m be significantly larger than n and that $[\lambda_i m]$ be used instead of $[\frac{im}{4}]$.

Simulation of this procedure was carried out for the quantiles of the Laplace distribution using $m = 16$, $i = 4, 8, 12$. The algorithm was started by using a single batch sample of size 16 to estimate θ_i, $i = 1, 2, 3$ and $1/2f(\theta_i)$. The results of this simulation are plotted in Figures 2 and 3. After the 120th iteration, $[A_p]_{ii}$, $i = 1, 2, 3$ are greater than the optimum values and close enough to ensure a near optimum convergence rate. Further, observe that X_{p_2}, the estimate of the median, appears to be biased slightly positive due to the random number of generator used in the simulation.

Joint Services Technical Advisory Committee
F44620-74-C-0056

P. Kersten and L. Kurz

REFERENCES

1. P. Kersten and L. Kurz, "Robustized Vector Robbins-Monro Algorithm with Applications to m-Interval Detection," Information Sciences, to appear.
2. G. Blom, "Statistical Estimates and Transformed Beta Variables," (New York: John Wiley and Sons, 1958).
3. Y.C. Ching and L. Kurz, "Nonparametric Detectors Based on m-Interval Partitioning," IEEE Trans. on Information Theory, IT-18, No. 2 (March 1972).
4. C. Derman and J. Sacks, "On Dvoretzky's Stochastic Approximation Theorem," Annals of Mathematical Statistics, 30, 601-606 (1959).
5. A. Dvoretzky, "On Stochastic Approximation," Third Berkeley Symposium, 1, 39-55 (1956).
6. J.G. Evans, "Robust Stochastic Approximation Procedures Using Nonparametric Statistics in Unspecified Noise Environments," Ph.D. Thesis, N.Y.U. (October 1973).
7. J.G. Evans, P. Kersten and L. Kurz, "Robust Recursive Estimation with Applications," to be published in Information Sciences.
8. J.L. Hodges and E.L. Lehmann, "Two Approximations to the Robbins-Monro Process," Third Berkeley Symposium, 1 (1956).
9. P. Lancaster, "Theory of Matrices," (New York: Academic Press, 1969).
10. J. Sacks, "Asymptotic Distribution of Stochastic Approximation Procedures," Annals of Mathematical Statistics, 29, 373-405 (1958).

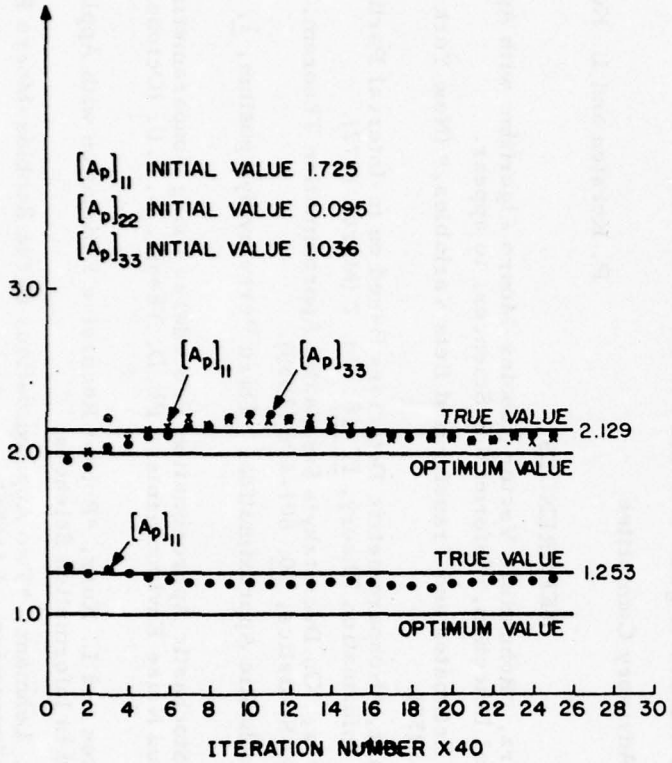


Fig. 2. Plot of adaptive gain coefficients using order statistics to estimate $1/2\,f(\theta_i)$, batch size 16.

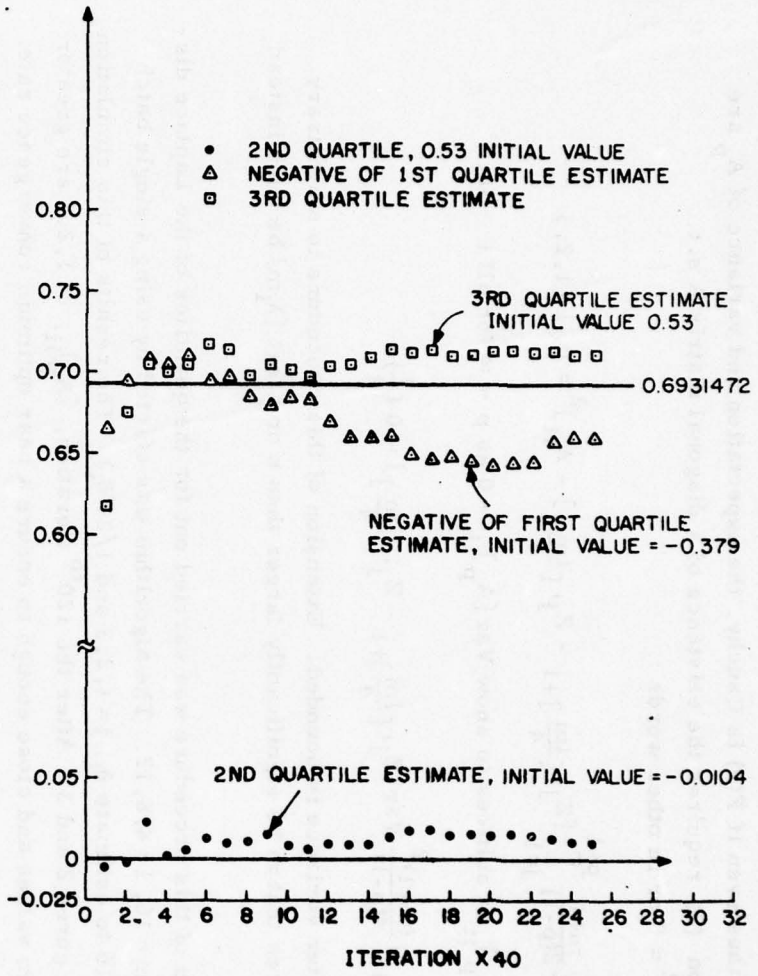


Fig. 3. Plot of Robbins-Monro procedure for quartiles of Laplace distribution using adaptive coefficients, batch size, N = 16.

M-LEVEL QUANTILE DETECTORS

B. Gotz and L. Kurz

The basic advantages of partition detectors based on quantile statistics are well established.[1-7] Some difficulties in simple implementation and operation of this type of detector in an adaptive mode were recently resolved for regular and mixture distribution noise environments.[8,9] The intent of this research is to extend the theory of partition detectors to M-level systems. Two types of M-ary detectors are considered -- a generalized-sign equiprobable partitioning detector (GSEPD) and a generalized-sign maximum-mean partitioning detector (GSMMPD). Unlike in the binary case, where the detector performance is evaluated on the basis of the asymptotic relative efficiency (ARE) in the Pitman-Noether sense, for evaluation of M-level detectors difference statistics with the associated efficacies, cross-efficacies and correlation coefficients are introduced.

A. Two Generalized Sign Detectors

Consider a simple generalization of the sign detector based on the test statistic

$$T = \frac{1}{L}\sum_{i=1}^{m} i\,n_i = \frac{1}{L}\sum_{j=1}^{L}\sum_{i=1}^{m} u(x_j - a_i) \tag{1}$$

where $\{x_j\}$ are L i.i.d. samples of observed data, $\{a_i\}$ are m quantiles of the distribution under the hypothesis (usually if noise only is present), $\{n_i\}$ are the m occupancy indicators of the number of samples between two adjacent quantiles or $\sum_{i=1}^{m} n_i = L$, and $u(\cdot)$ is the unit step function. It is easy to show that for the detector of Eq. (1), the efficacy for shift-of-the-mean alternatives reduces to

$$\delta_{S.M.} = \frac{\{\sum_{i=1} i[f(a_{i-1}) - f(a_i)]\}^2}{\sum_{i=1}^{m} i^2[F(a_i) - F(a_{i-1})] - \{\sum_{i=1}^{m} i[F(a_i) - F(a_{i-1})]\}^2} \tag{2}$$

where $F(\cdot)$ is the distribution under the hypothesis, $f(\cdot)$ is the corresponding probability density function (p.d.f.) and the quantiles $\{a_i\}$ are as yet unspecified. Two choices of quantiles result in two useful detectors:

(1) The choice of partitions is equiprobable or $F(a_i) - F(a_{i-1}) = \frac{1}{m}$ in yielding a generalized-sign equiprobable partitioning detector (GSEPD)

(2) The choice of the quantiles is such that the mean of the generalized sign statistic is maximized or $F(a_i) - F(a_{i-1}) = \frac{2i}{m(m+1)}$ and $F(a_i) = \frac{i(i+1)}{m(m+1)}$ yielding a generalized-sign maximum-mean partitioning detector (GSMMPD).

For GSEPD, Eq. (2) reduces to

$$\delta_{S.M.(1)} = \frac{12}{m^2-1}\left\{\sum_{i=1}^{m} i[f(a_{i-1}) - f(a_i)]\right\}^2 \tag{3}$$

If $m = 2$ and $f(a_0) = f(a_2) = 0$, GSEPD becomes a sign detector and $\delta_{S.M.(1)} = 4f^2(a_1)$ -- -- a well known result.

For the GSMMPD, Eq. (2) reduces to

$$\delta_{S.M.(2)} = \frac{18}{m^2+m-2}\left\{\sum_{i=1}^{m} i[f(a_{i-1}) - f(a_i)]\right\}^2 \tag{4}$$

Numerical calculations have been performed for both detectors under gaussian noise conditions. The ARE with respect to optimum detector of both detectors are over 90% for $m = 5$, with GSMMPD having a slight edge over the GSEPD. The ARE in non-gaussian noise with respect to the locally most powerful detector for gaussian noise frequently exceeds unity. As compared to rank detectors, both detectors have slightly reduced efficiency in gaussian noise but improved robustness if noise distributions vary. In addition, the two detectors are simple to implement and do not require expensive ranking of the observables.

B. Generalized Sign M-ary Detection

Unlike in the binary detection problem, a single partitioning based on the noise distribution is replaced by M partitionings, one "matched" to each hypothesis. For each hypothesis, H_i, $i = 1, \cdots, M$, a partitioning vector $\vec{a}_i$ and a statistic T_i must be selected. In general, for each hypothesis the test statistic is of the form

$$T_i = \frac{1}{L}\sum_{k=1}^{m} b_k n_{ik} \tag{5}$$

where the weight vector $\vec{b}$ of positive components is specified. An intuitively appealing choice of the partitioning vectors $\vec{a}_i$ is such that

$$E[T_i/H_i] = \vec{b} \cdot \vec{p}_i \tag{6}$$

are maximized, where $\vec{p}_i = (p_{i,1} \cdots p_{i,M})$ and $p_{i,j} = \text{Prob}[a_{i,j-1} < x \leq a_{i,j}]$. Since $\sum_{j=1}^{m} p_{i,j} = 1$, the optimum $\vec{a}_i$ satisfy

$$F_i(a_{i,j}) - F_i(a_{i,j-1}) = \frac{b_j}{\sum_{k=1}^{m} b_k}, \quad j = 1, \cdots, m \tag{7}$$

Since $F_i(a_{i,0}) = 0$, by iteration,

$$F_i(a_{i,j}) = \frac{\sum_{k=1}^{j} b_k}{\left[\sum_{k=1}^{m} b_k\right]} \tag{8}$$

are obtained.

Specifically, T_i, $i = 1, \cdots, M$ may be selected as GSMMP statistics considered in the previous subsection. In this case

$$T_i = \frac{1}{L} \sum_{k=1}^{m} k n_{ik} \tag{9}$$

where the quantile vectors are given by the solution of

$$F_i(a_{i,k}) = \frac{k(k+1)}{m(m+1)} \quad i = 1, \cdots, M \quad \text{and} \quad k = 1, \cdots, m \tag{10}$$

When the design of a detector is based on one hypothesis (e.g., the null hypothesis), as is the case in binary detection, the efficacy does not depend on the actual hypothesis separation but only on the local rate of separation. This is not a meaningful measure of performance when M "matched" detectors are used in an M-ary system. If we assume stochastically ordered hypotheses as representing the M-ary system, the major contribution of error results from a decision on a hypothesis adjacent to a true hypothesis. Thus, we need a measure of performance of two adjacent statistics, T_i and T_j. We replace the relative comparison to equivalent threshold comparison based on the difference statistic

$$T = T_i - T_j \tag{11}$$

The efficacy associated with the difference statistic is then a meaningful measure of performance of a "matched" type M-ary detector. It can be easily shown that the efficacy of T is

$$\delta_T = \lim_{L\to\infty} \frac{m_i^2 + m_j^2 - 2m_i m_j}{L(\sigma_i^2 + \sigma_j^2 - 2\rho\sigma_i\sigma_j)} = \lim_{L\to\infty} \left\{ \left[\frac{m_i^2}{L\sigma_i^2} + \frac{m_j^2}{L\sigma_j^2} - \frac{2m_i m_j}{L\sigma_i^2} \right]^{-1} \right.$$
$$\left. + \left[\frac{m_i^2}{L\sigma_j^2} + \frac{m_j^2}{L\sigma_j^2} - \frac{2m_i m_j}{L\sigma_j^2} \right]^{-1} - 2\rho \left[\frac{m_i^2}{L\sigma_i\sigma_j} + \frac{m_j^2}{L\sigma_i\sigma_j} - \frac{2m_i m_j}{L\sigma_i\sigma_j} \right]^{-1} \right\}^{-1} \tag{12}$$

where $m_k = \frac{\partial}{\partial\theta} E[T_k]\Big|_{\theta=0}$ k = i or j, θ - a signal parameter, and

$$\text{var } T = \sigma_i + \sigma_j^2 - 2\rho\sigma_i\sigma_j$$

with

$$\sigma_k^2 = \text{var } T_k,\ k = i \text{ or } j,\ \rho = E\{(T_j - \overline{T}_j)(T_i - \overline{T}_i)\}/\sigma_i\sigma_j$$

Introducing self- and cross-efficacies

$$\delta_{ii} = \lim_{L\to\infty} \frac{m_i^2}{L\sigma_i^2},\ \delta_{ij} = \lim_{L\to\infty} \frac{m_i^2}{L\sigma_j^2}$$

the efficacy of the difference statistic is

$$\delta_T = [(\delta_{ii} - \delta_{ji})^{-2} + (\delta_{ij} - \delta_{jj})^{-2} - 2\rho(\delta_{ii}\delta_{ij} + \delta_{ji}\delta_{jj} - \delta_{ii}\delta_{jj})^{-1}] \tag{13}$$

The self- and cross-efficacies in Eq. (13) are easy to calculate, The calculation of ρ in specific cases, though laborious as will be shown in the next section, does not create extreme difficulties. The difference efficacy of Eq. (13) may be used in the same fashion as transition probabilities in parametric detectors to evaluate the performance of M-ary partition detectors.

C. Efficacy of the Generalized Sign M-ary Detectors for the Class of Lehman Alternatives

Before the efficacy of Eq. (13) can be evaluated for the class Lehmann alternatives, the first and second moments of the statistic T_i of Eq. (5) must be calculated. Consider the statistic

$$T' = L\bar{T} = \sum_{i=1}^{m} b_i n_i \tag{14}$$

It can be shown that the characteristic function of T' is

$$\Phi_{T'}(v) = \left[\sum_{i=1}^{m} p_i e^{jb_i v}\right]^L \tag{15}$$

where p_i is defined in connection with Equation (6). Applying the change-of-scale theorem to Eq. (15) and using the method of characteristic functions, the mean and variance of $\bar{T}$, given by Eq. (14), are

$$E[\bar{T}] = \sum_{i=1}^{m} p_i b_i \tag{16}$$

and

$$\text{var}\,[\bar{T}] = \frac{1}{L}\left\{\sum_{i=1}^{m} p_i b_i^2 - [E(\bar{T})]^2\right\} \tag{17}$$

If in the difference statistic of Eq. (11) $T_i = T_1$ and $T_j = T_2$, for the class of Lehmann alternatives under the hypotheses H_i, i = 1 or 2

$$H_1 : F_1(x) = F(x) \quad \text{and} \quad F_2(x) = F^{1+\theta}(x) = e^{(1+\theta)\,\ell n\, F(x)}$$

where θ is the nonzero parameter representing the signal. For small θ and $b_j = j$, it follows from Eq. (16) that

$$m_i = E[T_i] = \sum_{j=1}^{m} j\left[F(a_{i,j})\,\ell n\, F(a_{i,j}) - F(a_{i,j-1})\,\ell n\, F(a_{i,j-1})\right] \quad i = 1 \text{ or } 2 \tag{18}$$

For i = 1, Eq. (18), after substitution from Eq. (10), reduces to

$$m_1 = \sum_{j=1}^{m} j\left[\frac{j(j+1)}{m(m+1)}\,\ell n\left(\frac{j(j+1)}{m(m+1)}\right) - \frac{j(j-1)}{m(m+1)}\,\ell n\left(\frac{j(j-1)}{m(m+1)}\right)\right]$$

which with the notation $c_j = \frac{j(j+1)}{m(m+1)}$ becomes

$$m_1 = \sum_{j=1}^{m} j(c_j\,\ell n\, c_j - c_{j-1}\,\ell n\, c_{j-1})$$

For i = 2, $F^{1+\theta}(a_{2,j}) = c_j$ or $F(a_{2,j}) = c_j^{\frac{1}{1+\theta}}$. Thus

$$m_2 = \sum_{j=1}^{m} j\left[c_j^{\frac{1}{1+\theta}}\,\ell n\, c_j^{\frac{1}{1+\theta}} - c_{j-1}^{\frac{1}{1+\theta}}\,\ell n\, c_j^{\frac{1}{1+\theta}}\right]$$

Similarly, the variances are

$$L\sigma_1^2 = \sum_{j=1}^{m} j^2\,\frac{2j}{m(m+1)} - \left[\sum_{j=1}^{m} j\,\frac{2j}{m(m+1)}\right]^2 = \frac{m^2+m-2}{18}$$

and

$$L\sigma_2^2 = \sum_{j=1}^{m} j\,2\left[c_j^{\frac{1}{1+\theta}} - c_{j-1}^{\frac{1}{1+\theta}}\right] - \left\{\sum_{j=1}^{m} j\left[c_j^{\frac{1}{1+\theta}} - c_{j-1}^{\frac{1}{1+\theta}}\right]\right\}^2$$

The efficacies and cross efficacies are then

$$\delta_{ij} = \frac{m_i^2}{L\sigma_j^2} \qquad \begin{matrix} i = 1,2 \\ j = 1,2 \end{matrix}$$

It remains to calculate the correlation coefficient

$$\rho = \frac{E[T_1 T_2] - E[T_1]\,E[T_2]}{\sigma_1 \sigma_2}$$

where all expectations are with respect to H_1, i.e., the reference distribution, $F(x)$. Since the variances have been calculated and the means are

$$E[T_1 / H_1] = \sum_{j=1}^{m} j p_{1,j} = \frac{2}{m(m+1)} \sum_{j=1}^{m} j^2 = \frac{2m+1}{3}$$

and

$$E[T_2 / H_1] = \sum_{j=1}^{m} j p_{2,j} = \sum_{j=1}^{m} j[F(a_{2,j}) - F(a_{2,j-1})] = \sum_{j=1}^{m} j\left[c_j^{\frac{1}{1+\theta}} - c_{j-1}^{\frac{1}{1+\theta}}\right]$$

It remains to calculate $E[T_1 T_2 / H_1]$. The cross-expectation is

$$E[T_1 T_2 / H_1] = \frac{1}{L^2} \sum_{i=1}^{m} \sum_{j=1}^{m} ij\, E[n_{1,i}\, n_{2,j} / H_1]$$

which after some mathematical manipulations reduces to

$$E[T_1 T_2 / H_1] = E[T_1 / H_1]\, E[T_2 / H_2] + \frac{1}{L} \sum_{i=1}^{m} \sum_{j=1}^{m} ij\, \text{Prob}[x \in A_{1,i} \cap A_{2,j}]$$

where

$$A_{k,\ell} = (a_{k,\ell-1},\ a_{k,\ell}]$$

To complete the evaluation of the correlation coefficient, an expression for $\text{Prob}[x \in A_{i,1} \cap A_{2,j}]$ must be found. These probabilities depend on the signal parameter θ in a highly nonlinear fashion. The important fact is that as θ increases from zero, the m partitioning intervals monotonically drift to the right (both ends of the interval drift to the right irrespective of what happens to the interval length). For small θ, $A_{1,i} \cap A_{2,j} = \phi$ except for $j = i$, $i = 1, \cdots, m$ and $j = i = 1$, $i = 2, \cdots, m$. It follows that

$$\text{Prob}[x \in A_{1,i} \cap A_{2,j}] =$$

$$= \begin{cases} F(a_{1,i}) - F(a_{2,i-1}) = c_i - c_{i-1}^{\frac{1}{1+\theta}} & \text{for } j = i;\ i = 1, \cdots, m \\ F(a_{2,i-1}) - F(a_{1,i-1}) = c_{i-1}^{\frac{1}{1+\theta}} - c_{i-1} & \text{for } j = i-1;\ i = 2, \cdots, m \\ 0 & \text{otherwise} \end{cases}$$

Thus, all components of the efficacy, δ_T, of Eq. (13) for Lehmann alternatives are calculated. It should be noted that there exists a range $\theta < \theta_{max}$ for which these results are valid; namely, θ_{max} corresponds to the point where $a_{2,i}$ crosses over to the right of $a_{1,i+1}$. It is easy to show that

$$\theta_{max} = \max_{2<i<m} \frac{\ell n\left[\frac{m(m+1)}{i(i-1)}\right]}{\ell n\left[\frac{m(m+1)}{i(i+1)}\right]} - 1 \quad i = 2, m-1$$

Joint Services Technical Advisory Committee
F44620-74-C-0056

B. Gotz and L. Kurz

REFERENCES

1. Y.C. Ching and L. Kurz, "Nonparametric Detectors Based on m-Interval Partitioning," IEEE Trans. on Information Theory, IT-18, No. 2, 251-257 (March 1972).
2. W.J. Bushnell and L. Kurz, "The Optimization and Performance of Detectors Based on Partition Tests," Proc. 1974 Conference on Circuits and Systems.
3. L. Kurz and G. Pasternack, "Nonparametric Classification Using Quadratic Quantile Statistics," 1976 IEEE Information Theory Symposium, Renneby, Sweden (June 21-24, 1976).
4. L. Kurz, "Rank and Nonrank Nonparametric Methodology in Classification Problems," 1973 International Symposium on Information Theory, Ashkelon, Israel (June 1973).
5. L. Kurz, "Nonparametric Detectors Based on Partition Tests," Supplementary Notes, EE704, PINY (September 1975).
6. W.J. Bushnell, "Optimization and Performance of Detectors Based on Partition Tests," Ph.D. Dissertation, PINY (June 1975).
7. G. Pasternack, "Nonparametric Classification Using Quadratic Quantile Statistics," Ph.D. Dissertation, PINY (June 1975).
8. P. Kersten and L. Kurz, "Bivariate m-Interval Detectors with Application to Edge Detection," Information and Control, to appear.
9. P. Kersten and L. Kurz, "Robustized Vector Robbins-Monro Algorithm with Applications to m-Interval Detection," Information Sciences, to appear.

SIGNAL SELECTION AND OPTIMUM SIGNAL DESIGN USING ROBUST PARTITION STATISTICS

P. Kersten and L. Kurz

The statistical properties of the m-Interval Partition Detectors and their variants are well developed. A concise exposition of the present state of the art may be found in Reference 2. In particular, the linear m-Interval Partition Test is nonparametric, robust, simple to implement, and in its most powerful form operates with little loss of efficiency in comparison to the locally most powerful detector. Moreover, in severe impulsive noise environments, the m-Interval Partition Detector retains its robustness. Bushnell[3] has considered the performance of the m-Interval Partition Detector (MIPD) in Gaussian and impulsive noise and compared it to the performance of the likelihood ratio detector designed for the Gaussian noise. The results of this study support the claim that the MIPD retains its robustness even if the noise distribution is replaced by the mixture model (Gaussian and impulsive noise).

In this report, the theory of the m-Interval Detector is extended to handle side constraints on non-constant signals and optimization of the associated detector. Using the derived locally most powerful scores, the performance of this detector is investigated. A discrete formulation is used to select the signal and correlation function of the detector to optimize a performance index which reflects the system constraints. Though the detector appears to be similar to a conventional correlation detector, it retains the robustness properties of the m-Interval Detector.

A. Problem Statement

Consider the system block diagrammed in Figure 1. Unlike the usual detector for non-constant signals, where one specifies the form of the signal and obtains the locally most powerful scores,[2] both the signal and the correlation function (which corresponds to d_k of Ref. 2) remain unspecified. The m-Interval Partition Test is expressed in the form (Although the form of T in Eq. (1) appears to be that of a linear matched filter, it possesses the robustness of the m-Interval Partition Statistic.)

$$T = \frac{1}{n} \sum_{k=1}^{n} g(k)\, h(X(k)) \tag{1}$$

where $X(k) = n(k) + s(k)$. Its relation to the form of the statistic used in Ref. 2 will be established later. The function $h(\cdot)$ is assumed to be specified and the i.i.d. noise samples are denoted by n(k) and the index k. The question is posed: Given a performance measure, what choice of signal s(k) and correlation function g(k) optimizes the performance measure of the system? This problem statement is one of the many forms of the Discrete Signal Design Problem, and is not complete until both the performance

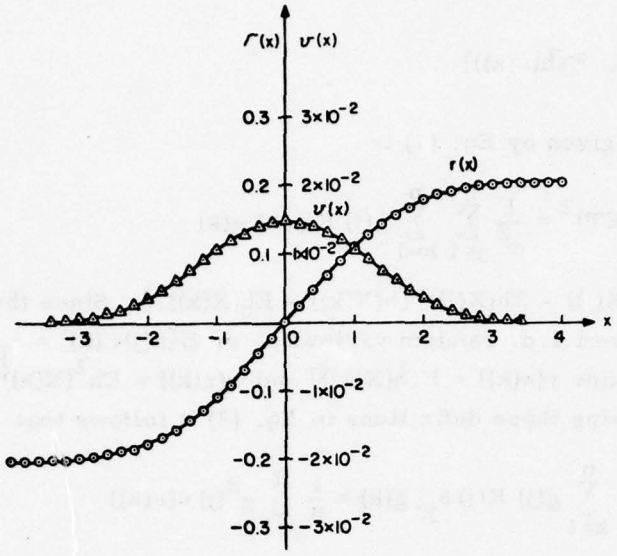


Fig. 1. Plot of r(x) and v(x) vs. x for the LMP scires and standard normal distribution using quantile 0.125, 0.25, 0.375, 0.5, 0.625, 0 0.750, 0.875

measure and the constraints imposed by the system are specified. In practice, the designer has to ensure that the problem statement represents a good approximation to the actual physical system and to the constraints that motivate the problem; on the other hand, he must also relax and/or tighten various conditions to either obtain non-trivial solutions or limit the number of solutions. The discrete signal approach as explained by Silver and Kurz,[4] has several advantages. In many systems, the detection scheme is actually applied to sampled data and so this approach better models many actual physical systems. The discrete version of the optimization problem is often more amenable to solutions. In addition, the increased usage of digital signal processing techniques in information processing systems further motivates the discrete formulation of this problem.

To relate the form of the m-Interval Partition Test given by Eq. (1) to the usual form, one equates the expected value of both of these forms to obtain

$$E[h\{s(k) + n(k)\}] = \sum_{\lambda=1}^{m} c_i [F(a_i - s(k)) - F(a_{i-1} - s(k))] = Eh(s(k))$$

and thus $d_k = g(k)$ where the notation of Ref. 2 is used. The expected value of T is then

$$E[T] = \frac{1}{n} \sum_{k=1}^{n} g(k)\, E[h(s(k))] \quad .$$

The variance of T as given by Eq. (1) is

$$\mathrm{Var}(T) = E(T - ET)^2 = \frac{1}{n^2} \sum_{j=1}^{n} \sum_{k=1}^{n} g(j)\, R(j,k)\, g(k) \tag{2}$$

where $R(j,k) = E\{[h(X(j)) - Eh(X(j))]\,[h(X(k)) - Eh(X(k))]\}$. Since the noise samples n(j) and n(k) are zero mean i.i.d. random variables, or $E[n(j)\,n(k)] = \delta_{jk}$, where δ_{jk} is the Kronecker delta. Define $r(s(k)) = E[h(X(k))]$ and $w(s(k)) = Eh^2(X(k))$ and $v(s(k)) = w(s(k)) - r^2(s(k))$. Using these definitions in Eq. (2) it follows that

$$n\mathrm{Var}(T) = \frac{1}{n} \sum_{j=1}^{n} \sum_{k=1}^{n} g(j)\, R(j)\, \delta_{jk}\, g(k) = \frac{1}{n} \sum_{j=1}^{n} g^2(j)\, v(s(k))$$

Nawrocky[5] has considered the design of optimal signals and detectors to minimize the effects of noise and intersymbol interference in the transmission of digital data using the discrete-time approach. This formulation allowed him to directly implement the results in simple digital structures. This particular formulation of the problem optimizes a cost functional which incorporates the signal-to-noise ratio and the Gabor bandwidth criterion.[6] Following Nawrocky, one seeks to maximize the functional

$$J' = \hat{h}A \sum_{k=o}^{K-1} g(k)\, r(s(k)) - \hat{h}B \sum_{k=o}^{K-1} g^2(k)\, v(s(k)) - \hat{h}C \sum_{k=o}^{K-1} (\Delta g(k))^2$$

$$- \hat{h}D \sum_{k=o}^{K-1} (\Delta s(k))^2 = S_1 - S_2 - S_3 - S_4 \tag{3}$$

where $\hat{h} = 1/K$ = the stepping increment and Δ is the forward difference operator. The first term S_1 is the expected value of the statistic T and the second term S_2 is $K\,\mathrm{Var}(T)$. The third term S_3 is a measure of the bandwidth of the correlation function and the fourth term S_4 is a measure of the bandwidth of the signal.

The functional J' is maximum when the last three terms of Eq. (3) are minimized and the first term is maximized subject to the constraints that will be specified later. Physically, this means the bandwidth of both the signal and the correlation function are to be minimized along with $K\,\mathrm{Var}(T)$ of the statistic. In contrast, the expected value of the statistic T is to be maximized. The s(k) and g(k) which optimize J' maximize the signal-to-noise ratio of the system while reducing the bandwidth of both s(k) and

g(k). The definition of both r(s(k) and v(s(k)) involve the CDF of the underlying noise distribution and therefore the signal problem is dependent upon the underlying noise distribution as well as the scores of the statistic T. Accordingly, assume $\Phi(x)$ is the standard normal CDF and c_i are the locally most powerful scores. Bushnell[3] has shown that with Gaussian scores that the linear m-Interval Partition Test performs well even in Gaussian and Impulsive noise environments. Thus Gaussian scores are a good initial approximation for studying this optimization problem since the performance of the statistic is relatively insensitive to perturbations in the Gaussian assumption. Figure 1 contains plots of both r(x) and v(x) for this case. Note that the first two derivatives of both r and v are continuous and bounded, i.e., $r, v \in C^2$. Since $\hat{h}$ is common to all the terms in J', $J = J'/\hat{h}$ is maximized instead of J'. To obtain a physically meaningful solution, energy constraints must be placed upon both g(k) and s(k). That is both

$$\sum_{k=o}^{K-1} s^2(k)\hat{h} \quad \text{and} \quad \sum_{k=o}^{K-1} g^2(k)\hat{h}$$

are required to be equal to given constants. Couching the above in the terminology of optimal control, allows application of all the optimization techniques developed for control theory to the signal design problem. Define $x_1(k) = s(k)$ and let

$$u_1(k) = \Delta x_1(k) \quad i.e., \quad x_1(k+1) = x_1(k) + \hat{h}\, u_1(k) \tag{4}$$

Thus $u_1^2(k)$ is a measure of the Gabor bandwidth of s(k). To express the energy constraint upon s(k), let

$$\Delta x_2(k) = x_1^2(k) \quad \text{or} \quad x_2(k+1) = x_2(k) + \hat{h}\, x_1^2(k) \tag{5}$$

so that

$$\sum_{k=1}^{K-1} \Delta x_2(k) = \sum_{k=1}^{K-1} s^2(k) = \text{power of the signal } s(k).$$

In like manner define $x_3(k) = g(k)$ and let

$$u_2(k) = \Delta x_3(k) \quad \text{or} \quad x_3(k+1) = x_3(k) + \hat{h}\, u_2(k) \tag{6}$$

so that $u_2^2(k)$ is a measure of the Gabor bandwidth of g(k). In addition,

$$\Delta x_4(k) = x_3^2(k) \quad \text{or} \quad x_4(k+1) = x_4(k) + \hat{h}\, x_3^2(k) \tag{7}$$

implement the energy constraint upon $x_4(k)$.

Passing to vector notation define $\underline{x}(k)^T = (x_1(k), x_2(k), x_3(k), x_4(k))$, as the state vector and $\underline{u}(k)^T = (u_1(k), u_2(k))$ as the control vector of the system defined by Equations (4) to (7). The initial and final values of the state vector are then given by $x(0) = \underline{0}$ and $\underline{x}(K)^T = (0, E_s, 0, E_g)$ respectively. J is then expressed by

$$J = \sum_{k=o}^{K-1} L(\underline{x}(k), \underline{u}(k)) \tag{8}$$

where $L(\underline{x}(k), \underline{u}(k)) = A x_3(k) r(x_1(k)) - B x_3^2(k) v(x_1(k)) - C u_2^2(k) - D u_1^2(k)$ and the system equations can be expressed in a generic form as

$$\underline{x}(k+1) = \underline{f}(\underline{x}(k), \underline{u}(k)) \tag{9}$$

The above formulation permits the direct application of the Gradient Projection Algorithm. In fact, the algorithm described by Dyer and McReynolds for implementing this method is directly applicable provided the vector of parameters $\underline{a}$ is set equal to zero (Ref. 7, pp. 57-61). This algorithm implements the gradient projection method by iteratively calculating the proper gradients with respect to the control variables and thus producing an efficient computer program. This approach is considered in Section B.

B. The Gradient Projection Method for Solution of the Signal Design Problem Using the m-Interval Partition Statistic

The Gradient Projection Algorithm (GPA) is simple in concept although complex in analysis and application. This method optimizes a functional subject to constraints by an iterative procedure which is a variation of the steepest ascent (descent) method. This algorithm replaces the gradient of the functional J with a projected gradient where the projection is upon the feasible region.[8]

The implementation of this method by Dyer and McReynolds[7] employs a format which is amenable to digital computation. Their approach is to treat the gradient projection method via the method of Lagrange Multipliers using the constraints to form an augmented functional. To be specific, let $L(\underline{a})$ represent the function which one wishes to maximize and let $\underline{M}(\underline{a}) = \underline{0}$ be a vector constraint to be satisfied by any feasible point. One constructs the augmented functional

$$J^* = L + (\underline{M} - \underline{Q})^T \underline{\nu}^k \tag{10}$$

where $\underline{Q}$ is a vector representing the constraint level (normally taken as zero in the standard Lagrange formulation) and $\underline{\nu}^k$ represents a vector of multipliers at the k-th iteration. One then proceeds in obtaining the gradient correction to a feasible solution $\underline{a}^k$ at the k-th step in precisely the same manner as with the steepest ascent method.

This yields,

$$\delta \underline{a} = \underline{a}^{k+1} - \underline{a}^{k} = \varepsilon(\underline{L}_a^T + \underline{M}_a^T \underline{\nu}^k) \tag{11}$$

where ε = stepping parameter and $\underline{L}_a$ is the partial derivative of L with respect to the vector of parameters $\underline{a}$ which is a n x 1 vector since Dyer and McReynolds define $\underline{L}_a = (\partial L/\partial a_1, \cdots, \partial L/\partial a_n)$. Following their notation, $\underline{M}_a$ is a matrix $[\partial M_i/\partial a_j]$, $j = 1, \cdots, n$; $i = 1, \cdots, q$. The vector $\underline{\nu}^k$ is then calculated to ensure a predetermined small change in the constraint level $\underline{Q}$. That is, $\delta \underline{Q}$ is proportional to the error in the constraint $\underline{M}(\underline{a}) = \underline{0}$. Neglecting higher than linear terms in approximation of the functional, $\delta \underline{Q} = \underline{M}_a \, \delta \underline{a}$, and substituting $\delta \underline{a}$ from Eq. (11) into this equation yields

$$\delta \underline{Q} = \varepsilon(\underline{M}_a \underline{L}_a^T + \underline{M}_a \underline{M}_a^T \underline{\nu}^k)\Big|_{\underline{a} = \underline{a}^k}$$

which when solved for $\underline{\nu}^k$ gives

$$\underline{\nu}^k = (\underline{M}_a \underline{M}_a^T)^{-1}(\delta \underline{Q}/\varepsilon - \underline{M}_a \underline{L}_a^T) \, . \tag{12}$$

Thus to implement this method, one uses the recursion

$$\underline{a}^{k+1} = \underline{a}^k + \varepsilon^k(\underline{L}_a + (\underline{\nu}^k)^T \underline{M}_a)^T \tag{13}$$

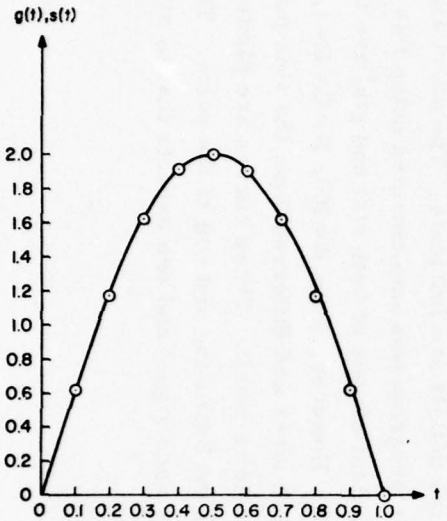
where $\underline{\nu}^k$ is given by Equation (12).

Application of this procedure to the signal design problem under consideration is straightforward in principle although the details are intricate. The exact algorithm is contained in Reference 7. A computer program was implemented using this algorithm and run with A = B = C = D = 1. The resultant shapes of both s(k) and g(k) are that of a sinusoidal pulse as plotted in Figure 2. However, when A = 100, B = C = D = 1, the curve for g(k) and s(k) are different from each other and different from the sine pulse, although the perturbation from the sinusoid is small. These curves are plotted in Figure 4. Note that s(k) has steeper sides at the beginning and end of the pulse. This is to be expected since r(s(k)) is near zero for zero signal and one expects that in order to increase

$$\sum_{k=o}^{K-1} g(k)\, r(s(k))$$

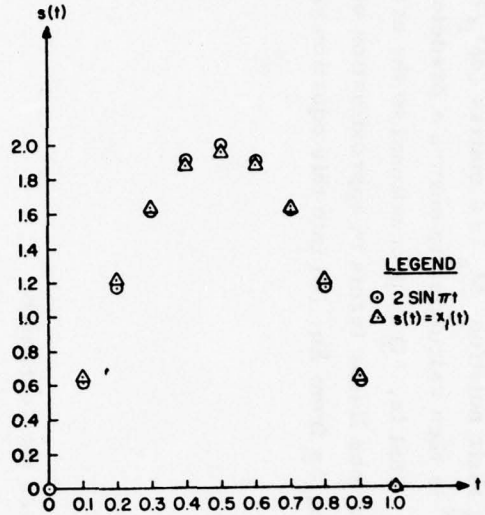
that s(k) will have to increase faster at the beginning and decrease faster at the end, but remain flatter in the center. This allows more contributions to

$$\sum_{k=o}^{K-1} g(k)\, r(s(k))$$



t	g(t)	2 SINπt
0.1	0.6184	0.618033
0.2	1.176	1.175572
0.3	1.618	1.618034
0.4	1.902	1.902113
0.5	2.000	2.00000
0.6	1.902	1.902113
0.7	1.618	1.618034
0.8	1.176	1.175572
0.9	0.6184	0.618033
1.0	1.788×10^{-7}	0.00000

Fig. 2. Plot of g(t) and s(t) vs. t for A=B=C=D=1, h=0.1 and LMP scores. Plot is 2 sin 2t to several decimal places, N=25 iterations of the gradient projection algorithm.



t	2 SINπt	s(t)	g(t)
0.1	0.618033	0.6492	0.6317
0.2	1.175572	1.210	1.191
0.3	1.618034	1.627	1.622
0.4	1.902113	1.880	1.893
0.5	2.0000	1.965	1.984
0.6	1.902113	1.880	1.893
0.7	1.618034	1.627	1.622
0.8	1.175572	1.210	1.191
0.9	0.618033	0.6492	0.6317
1.0	0.00000	1.788×10^{-7}	-1.788×10^{-7}

NOTE: g(t) IS NOT PLOTTED SINCE IT IS LOCATED BETWEEN s(t) AND 2 SINπt AS REFERENCE TO ACCOMPANYING TABLE WILL CONFIRM.

Fig. 3. Plot of s(t) and 2 sin πt vs. t for A=100, B=C=D=1, h=0.1 and standard normal and LMP scores. Gradient projection method.

from the initial and final values of the summation while still satisfying the energy constraints. In essence, these results imply that the signal design is not sensitive to the relative importance of the constraints and the signal-to-noise ratio, i.e., the types of constraints are the important factors in the functional for determining the shape of s(k) and g(k). This algorithm converges quickly (10-15 iterations) since $u_1(k)$ and $u_2(k)$ are known to be essentially cosines and are so chosen to initiate the program. The initial starting values were obtained from an application of the discrete maximum principle to a simplified version of the same problem. Though this problem was analyzed for Gaussian additive noise, the method of solution is applicable to other noise distributions possessing a continuous p.d.f. The procedure and the associated software are sufficiently general and flexible to include other constraints without any difficulty.

Joint Services Technical Advisory Committee
F44620-74-C-0056

P. Kersten and L. Kurz

REFERENCES

1. Y.C. Ching and L. Kurz, "Nonparametric Detectors Based on m-Interval Partitioning," IEEE Trans. on Information Theory, IT-18, No. 2, 251-257 (March 1972).
2. L. Kurz, "Nonparametric Detectors Based on Partition Tests," Supplementary Notes, EE704, PINY (September 1975).
3. W.J. Bushnell, "Optimization and Performance of Detectors Based on Partition Tests," Ph.D. Dissertation, Polytech. Inst. of New York (1975).
4. H.I. Silver and L. Kurz, "A Class of Discrete Signal Design Problems in Burst Noise," IEEE Trans. on Information Theory, IT-18, No. 2 (March 1972).
5. R.J. Nawrocky, "A Discrete Time Approach to Signal and Detector Design in Additive Gaussian and Burst Noise and Intersymbol Interferences," Ph.D. Dissertation, Polytech. Inst. of New York (1975) and R.J. Nawrocky and L. Kurz, "Design of Digital Filters for Extraction of Signals for Inter-Symbol Interference and Noise," 8th Annual Conference on Information Sciences and Systems, Princeton, N.J. (March 28-29, 1974).
6. D. Gabor, "Theory of Communications," J. Inst. Electrical Engineers (London), 93, Part 3, 427-457 (November 1946).
7. P. Dyer and S.R. McReynolds, "The Computation and Theory of Optimal Control," Academic Press, N.Y. (1970).
8. J.B. Rosen, "The Gradient Projection Method for Non-Linear Programming. Part I. Linear Constraints," J. of Soc. for Industrial and Applied Math., 8, No. 15 (March 1960).
9. P. Kersten, "Robustized Recursive Estimation and Adaptive Partition Detectors," Ph.D. Dissertation, Polytech. Inst. of New York (1976).

A NOVEL FM DETECTOR FOR SUPPRESSION OF INTERCHANNEL INTERFERENCE

F.A. Cassara and H. Schachter

Numerous investigators[1-5] have focused their attention on the problem of interchannel interference in FM receivers. Some studies specialize to analog FM; others to digital frequency or phase shift keyed signals. All studies conclude that the presence of interfering signals degrade the quality of reception severely.

Interfering signals can arise in the following general areas:

(1) Co-channel interferer, e.g., (a) neighboring transmitters sharing the same frequency band; (b) spurious signals such as an image channel in a superheterodyne receiver; (c) multipath echoes

(2) Adjacent channel interferer, e.g., (a) inadequate selectivity in the receiver's IF filter; (b) "crowding" in the radio frequency spectrum; (c) spurious signals.

In this report a novel FM detector which has demonstrated capability in suppressing the degradation in receiver performance due to the presence of an interfering signal is presented.

A. The Novel FM Detector

The block diagram of the novel FM detector is shown in Figure 1. The principle

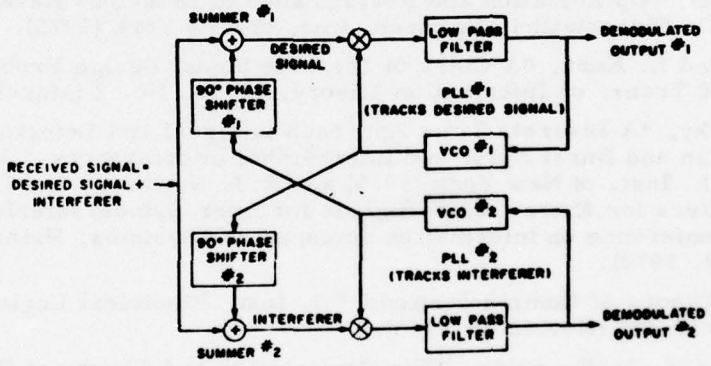


Fig. 1. Novel FM detector for suppression of interchannel interference.

of operation can be described as follows: assume the input signal consists of a frequency modulated carrier plus a frequency modulated interferer. The interferer to desired carrier amplitude ratio is denoted as η. Phase-locked loop (PLL) #1 locks on to and tracks (by the capture effect) the stronger of the two received FM signals but its voltage controlled oscillator output signal (VCO #1) lags by approximately 90°. An additional 90° phase shift is introduced by phase shifter #2 so that the signal appearing at phase shifter #2's output is 180° out of phase with respect to the stronger received signal.

By proper adjustment of the gain constants of summer #2 the stronger received signal is cancelled leaving only the weaker of the two received FM signals at the input to PLL #2. The instantaneous phase of VCO #2 output signal tracks the instantaneous phase of the weaker signal but lags by 90°. An additional 90° phase shift is introduced by phase shifter #1 producing a signal at the output of phase shifter #1 which is 180° out of phase with respect to the weaker of the two received FM signals. The weaker signal can thus be cancelled at summer #1 leaving only the stronger signal appearing at the input to PLL #1. Since this novel detector has two separate outputs -- namely the outputs of the individual phase-locked loops it possesses the capability of demodulating both the stronger and the weaker received signals even though they may be co-channel and share the same frequency band. This is a task impossible with any of the other existing FM detectors since they all obey the well known capture effect.

B. Experimental Results

Hardware for this new detector was constructed using monolithic PLL integrated circuits (Signetics 562B) to detect square wave FM carriers. The novel device successfully demodulated both the stronger and weaker sinusoidally modulated FM carriers for the case where the stronger signal was as much as 20 dB greater than the weaker carrier. Figure 2 describes the response of the novel demodulator and limiter-

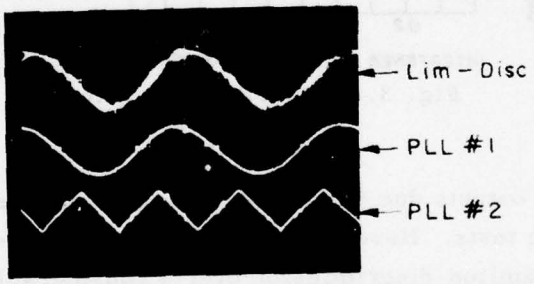


Fig. 2. Response of Lim-Disc. and novel FM detector with $\eta = 0.5$.

discriminator (General Radio 1142-A) when simultaneously driven by a 455 kHz square wave carrier plus a co-channel 455 kHz FM square wave interferer. The desired carrier was frequency modulated by a 100 Hz sinusoid while the interferer was frequency modulated by a 200 Hz triangle wave. The peak frequency deviation of both carriers was 7 kHz. The interferer to desired carrier amplitude ratio was set at $\eta = 1/2$. Each output was filtered by identical 6 kHz bandwidth post detection low pass filters (Krohn-Hite Model 3202). Each PLL loop bandwidth was designed to be 100 kHz (much wider than the 7 kHz peak frequency deviation of the input carriers). In addition each PLL was essentially a first order loop with a wide (20 kHz) bandwidth single pole RC low

pass loop filter. The results of Fig. 2 reveal that the novel detector can demodulate both the stronger and weaker FM carriers with considerable improvement over the limiter-discriminator which, by the capture effect, demodulates only the stronger received FM signal.

More quantitative results related to this oscillogram are shown in Fig. 3 where the detected output SNR vs. η is plotted. Noise here is interpreted as distortion ap-

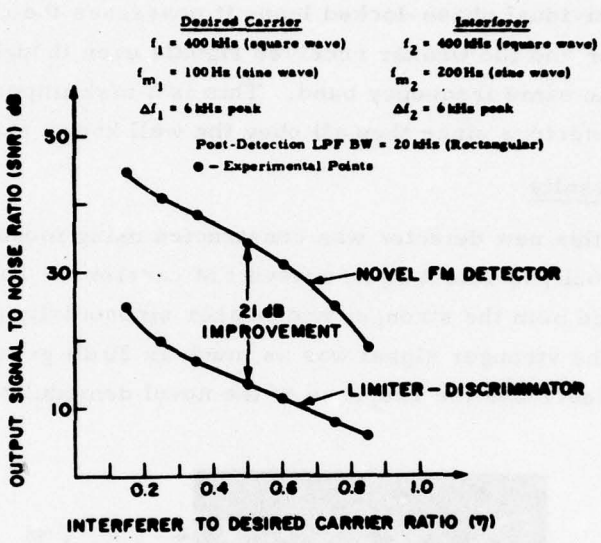


Fig. 3. Output SNR vs. η for Lim-Disc. and novel FM detector.

pearing in the detected outputs due to the presence of an interferer. No random noise was introduced in these tests. Here we see that the novel detector offers 21 dB improvement compared to the limited-discriminator over a considerable range of η.

The sensitivity of the novel detector of Fig. 1 to amplitude fluctuations in the received signals was experimentally evaluated by driving the demodulator with a 10 volt peak-to-peak 457 kHz square wave carrier frequency modulated by a 100 Hz triangle wave (see upper trace of oscillogram in Fig. 4) plus a co-channel 457 kHz FM square wave interferer amplitude modulated by a 50 Hz sine wave and simultaneously freq. mod. by a 200 Hz sinusoidal signal (second trace in Figure 4). Both FM carriers had a 10 kHz peak-to-peak frequency deviation. The third and fourth traces illustrate the demodulated outputs of PLL #1 and PLL #2, respectively. Both PLL outputs were filtered by identical 10 kHz bandwidth low pass filters. As can be seen from the oscillogram the weaker FM carrier (second trace) varies in amplitude from a minimum of

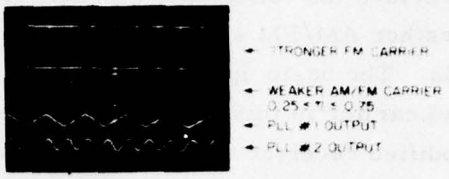


Fig. 4. Response of novel FM detector of Fig. 1 to AM/FM interferer.

2.5 volts peak-to-peak to a maximum of 7.5 volts peak-to-peak due to the 50 Hz sinusoidal amplitude modulation. This results in a variation in the FM carrier amplitude ratio, η, of $0.25 \leq \eta \leq 0.75$. Throughout this range the novel detector demodulated both co-channel FM signals quite well.

Figure 5 illustrates the block diagram of a modified receiver structure specifically designed to accommodate the case when the applied FM signals have amplitudes which

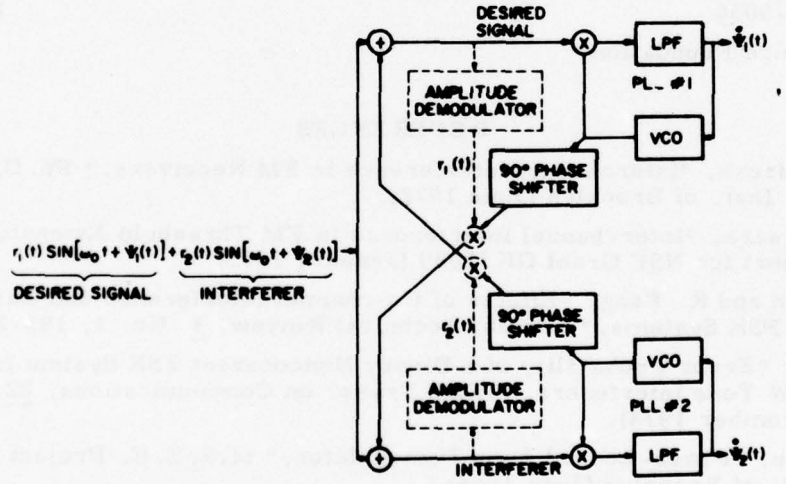


Fig. 5. Novel FM detector modified to demodulate AM/FM input signals.

vary appreciably such as occur in a "fading" interferer or a multiple interferer signal environment. In the latter case all the interferers may be "lumped" together analytically and represented as one AM/FM interferer. Modifications to the novel detector of

Fig. 1 are shown in dashed lines. The input signals are amplitude modulated by $r_1(t)$ and $r_2(t)$ and simultaneously phase modulated by $\psi_1(t)$ and $\psi_2(t)$ as shown in Figure 5. In this modified receiver structure the cancellation signals fed back are no longer constant amplitude signals but rather AM/FM signals and thus are capable of cancelling their respective input signals. The basic operating principle for this modified structure is the same as that presented earlier in this report. No experimental studies have thus far been obtained on this modified receiver structure and is left as a study for future research.

C. Conclusions

A novel FM detector capable of demodulating an FM signal corrupted by a co-channel interferer for the case when the interferer is as much as 20 dB stronger than the desired FM signal has been presented. Experimental results demonstrating such capability have been included. Such results revealed, among other things, that the demodulator was relatively insensitive to variations in the interferer to desired carrier amplitude ratio η. Future studies on the novel detector will include the effects of the simultaneous action of random noise and an interferer, multiple and fading interferers, and digital frequency shift keyed signals.

Joint Services Technical Advisory Committee
F44620-74-C-0056

F.A. Cassara

National Science Foundation
GK32689

REFERENCES

1. T.S. Sundresh, "Interchannel Interference in FM Receivers," Ph.D. Dissertation, Polytech. Inst. of Brooklyn (June 1972).
2. F.A. Cassara, "Interchannel Interference in FM Threshold Extension Receivers," Final Report for NSF Grant GK32689 (January 1975).
3. O. Shimbo and R. Fang, "Effects of Co-channel Interference and Gaussian Noise in M-ary PSK Systems," Comsat Technical Review, 3, No. 1, 183-206 (1973).
4. L. Wang, "Error Probability of a Binary Noncoherent FSK System in the Presence of Two CW Tone Interferers," IEEE Trans. on Communications, 22, No. 12, 1948-1949 (December 1974).
5. Gottesman, "Phase-Locked Loop Demodulator," M.S.E.E. Project Report, Polytech. Inst. of Brooklyn (June 1964).

OPTIMIZATION OF M-ARY DIGITAL TRANSMISSION SYSTEMS IN THE PRESENCE OF INTERSYMBOL INTERFERENCE AND NOISE

L. Kurz and R.J. Nawrocky

In this report, the problem of optimum signal and detector selection in the presence of intersymbol interference and noise as formulated in Refs. 1 and 2 is extended to M-ary transmission systems, where the signal set consists of M distinct waveforms. Tutorial treatments of M-ary communication problems may be found in Helstrom[3] and Weber.[4] The problem of generating M-ary alphabets with useful properties in burst noise has been considered by Silver and Kurz[5] using amplitude averaging techniques, while other aspects of the M-ary problem have been studied by Nuttall,[6] Nuttall and Amoroso,[7] and Nuttall and Floyd.[8]

In this report, the posed optimization problem is approached through the minimization of a general system cost functional based on the noise-to-signal ratio criterion and a cost matrix relating the properties of the individual signals and detectors. In particular, two forms of the cost matrix are considered leading to either the minimum probability of error or the maximum probability of detection solutions.

For simplicity of the presentation, the problem is initially developed for the all-pass channel and white Gaussian noise. Subsequently, the formulation is extended to include a bandlimited channel. The performance of M-ary systems in additive Gaussian noise and intersymbol interference is examined and compared on the basis of a constant data transmission rate.

A. Development and Minimization of the Cost Functional: All-pass Channel

Consider the general M-ary transmission system of Fig. 1 for communication over a single channel in additive Gaussian noise environment. For simplicity of the presentation, the channel is assumed to be all-pass. An extension to bandlimited channels is treated in the following section.

In the all-pass case, the generalized cost functionals for a signal $s_m(k)$ and a detector $g_m(k)$, each subject to the Gabor bandwidth constraints and $m = 1, 2, \cdots, M$, may be written using discrete formulation as

$$J_{sm} = \Delta \sum_{k=0}^{K-1} \left\{ \sum_{n=1}^{M} s_m(k)\, a_{mn}\, g_n(k) + \sum_{\ell=1}^{L_s} c_{\ell m} [\nabla^{(\ell)} s_m(k)]^2 \right\}$$

$$= \Delta \sum_{k=0}^{K-1} G_{sm}(k) \qquad (1)$$

and

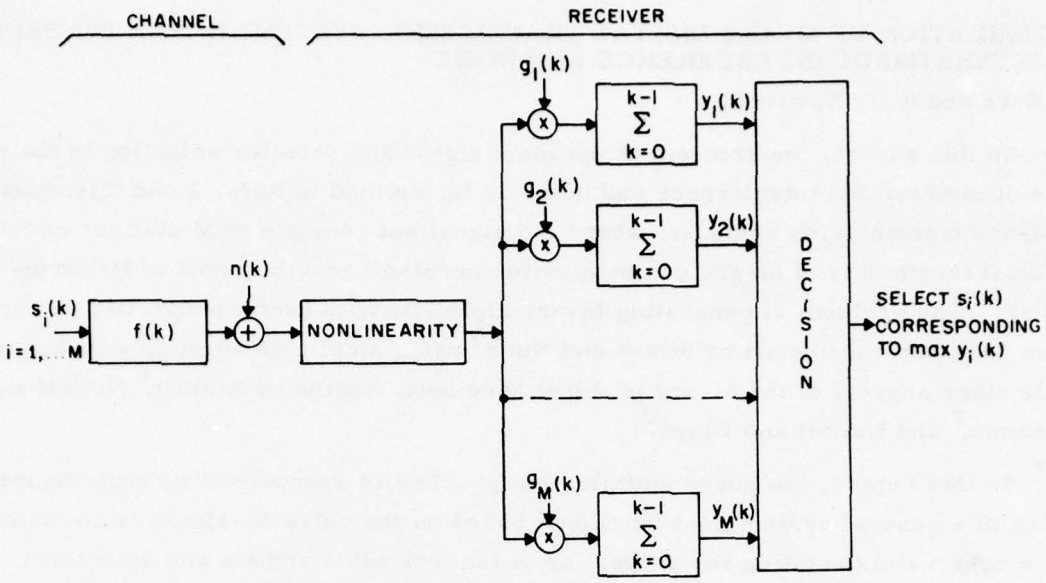


Fig. 1. A model of an M-ary transmission system.

$$J_{gm} = \Delta \sum_{k=0}^{K-1} \left\{ \sum_{n=1}^{M} s_n(k)\, a_{mn}\, g_m(k) + B g_m(k) \sum_{j=1}^{K-1} R(k,j)\, g_m(j) + \right.$$

$$\left. + \sum_{\ell=1}^{L_g} d_{\ell m} [\nabla^{(\ell)} g_m(k)]^2 \right\} = \Delta \sum_{k=0}^{K-1} G_{gm}(k) \tag{2}$$

respectively. In the above expressions, the coefficients a_{mn} represent the cost associated with the correlation between the m-th signal and the n-th detector. The remaining notation is similar to the notation of Reference 1. The overall system cost functional can be expressed in terms of Eqs. (1) and (2) as

$$J = \sum_{m=1}^{M} [J_{sm} + J_{gm}] \tag{3}$$

The correlations between pairs of signals and pairs of detectors may be included as side constraints

$$\nabla x_{smn}(k) = s_m(k)\, s_n(k) \tag{4}$$

and

$$\nabla x_{gmn}(k) = g_m(k)\, g_n(k) \tag{5}$$

for a total of $2M^2$ constraints. Additional $(L_s + L_g)\, M$ constraints arise from the bandwidth terms in Equations (1) and (2). For $L_s = L_g = 1$, these constraints become

$$\nabla s_m(k) = u_m(k) \tag{6}$$

and

$$\nabla g_m(k) = v_m(k) \tag{7}$$

The Hamiltonian corresponding to a signal-detector pair can be expressed in terms of Eqs. (1) through (7) and adjoint variables ψ_{mn}, ϕ_{mn}, p_m and q_m as

$$H_m(k) = -\Delta[G_{sm}(k) + G_{gm}(k)] + \sum_{n=1}^{M} [s_m(k)\,\psi_{mn}\,s_n(k) + g_m(k)\,\phi_{mn}\,g_n(k)] +$$

$$+ p_m(k+1)\,[s_m(k) + \Delta u_m(k)] + q_m(k+1)\,[g_m(k) + \Delta v_m(k)] \tag{8}$$

The variables ψ_{mn} and ϕ_{mn}, which can be shown to be time-independent, take on a special significance in this formulation because they represent the signal and the detector correlations coefficients, respectively. The overall Hamiltonian expressed in terms of Eq. (8) becomes

$$H(k) = \sum_{m=1}^{M} H_m(k) \tag{9}$$

Maximizing Eq. (9) with respect to the control variables $u_m(k)$ and $v_m(k)$, yields

$$u_m(k) = \frac{1}{2\,c_{1m}}\,p_m(k+1) \quad , \quad c_{1m} \geq 0 \tag{10}$$

$$v_m(k) = \frac{1}{2\,d_{1m}}\,q_m(k+1) \quad , \quad d_{1m} \geq 0 \tag{11}$$

which, in connection with Eqs. (6) and (7), gives

$$s_m(k+1) = s_m(k) + \frac{\Delta}{2\,c_{1m}}\,p_m(k+1) \tag{12}$$

$$g_m(k+1) = g_m(k) + \frac{\Delta}{2\,d_{1m}}\,q_m(k+1) \tag{13}$$

The adjoint equations are

$$p_m(k+1) = -\psi_{mn}\,s_m(k) + \Delta\,a_{mn}\,g_m(k) - \sum_{n=1}^{M} [\psi_{mn}\,s_n(k) - \Delta\,a_{mn}\,g_n(k)] + p_m(k) \tag{14}$$

$$q(k+1) = -\phi_{mn}\,g_m(k) + \Delta\,a_{mn}\,s_m(k) - \sum_{n=1}^{M} [\psi_{mn}\,g_n(k) - \Delta\,a_{mn}\,s_n(k)] +$$

$$+ 2\,B\,\Delta^2 \sum_{j=1}^{K-1} R(k,j)\,g_m(k) + q_m(k) \tag{15}$$

so that the optimal set of equations for a signal-detector pair consists of Equations (12) through (15). The minimization of the overall system cost function in Eq. (3) subject to correlation and first order bandwidth constraints on all signals and detectors, therefore, involves the solution of a system of 4M coupled, first order, linear sum-difference equations.

In view of the fact that the cost functional is expressed in terms of various correlation coefficients, the values of these coefficients must be specified for a unique solution of the optimization problem. The three types of coefficients may be considered to be the elements of three matrices: a signal-detector cross-correlation (cost) matrix $A = \{a_{mn}\}$, a signal correlation matrix $\Psi = \{\psi_{mn}\}$ and a detector correlation matrix $\Phi = \{\phi_{mn}\}$. While the matrix A may take on a number of forms, two forms are particularly interesting: those leading to formulations yielding the minimum probability of error, P_e, and those which maximize the probability of detection, P_d. The first form of the A matrix is given by (see Ref. 4, Chapter 10)

$$a_{mn} = 1 - \delta_{mn} \tag{16}$$

where δ_{mn} is the Kronecker delta, and the other by

$$a_{mn} = -\delta_{mn} \tag{17}$$

In Eq. (17) the negative sign arises from the fact that P_d is maximized. In terms of the matrices of Eqs. (16) and (17), the optimal equations, Eqs. (14) and (15), reduce to

$$p_m(k-1) = -\psi_{mn}\, s_m(k) - \Delta g_m(k) - \sum_{n=1}^{M} [\psi_{mn}\, s_n(k) - \Delta g_n(k)] + p_m(k) \tag{18}$$

$$q_m(k+1) = -\phi_{mn}\, g_m(k) - \Delta s_m(k) - \sum_{n=1}^{M} [\phi_{mn}\, g_n(k) - \Delta s_n(k)] + \\ + 2B\Delta^2 \sum_{j=1}^{K-1} R(k,j)\, g_m(k) + q_m(k) \tag{19}$$

for minimum of P_e, and to

$$p_m(k+1) = -\psi_{mn}\, s_m(k) - \Delta g_m(k) - \sum_{n=1}^{M} \psi_{mn}\, s_n(k) + p_m(k) \tag{20}$$

$$q_m(k+1) = -\phi_{mn}\, g_m(k) - \Delta s_m(k) - \sum_{n=1}^{M} \phi_{mn}\, g_n(k) + 2B\Delta^2 \sum_{j=1}^{K-1} R(k,j)\, g_m(k) + q_m(k) \tag{21}$$

for maximum of P_d. It should be noted that for the above two cases, the optimal equations differ in the sense that in the minimum of P_e the m-th adjoint variables are cross-coupled to the n-th signal or correlation function, while for the maximum of P_d no such coupling exists.

Similar to the matrix A, the correlation matrices Ψ and Φ may assume a variety of forms some of which, however, may lead to nonlinear boundary conditions. In the special case of orthogonal signals and detectors, these become

$$\psi_{mn} = \psi \delta_{mn} \tag{22}$$

and

$$\phi_{mn} = \phi \delta_{mn} \tag{23}$$

B. Extension to Bandlimited Channels

The formulation of the previous section can be directly extended to the case of a general bandlimited channel. However, for simplicity of the presentation, the problem is developed in detail for the specific case of a one-pole channel and white Gaussian noise environment. The problem is formulated for the maximum of P_d solution in terms of a set of orthogonal channel output signals. The solutions are considered for equal energy input signals using appropriate sets of parameter values and boundary conditions.

For a one-pole channel, the A matrix of the form of Eq. (17) and correlation matrices in Eqs. (22) and (23), the functional of Eq. (2) for $L_s = 2$ and $L_g = 0$ is

$$J = \sum_{m=1}^{M} \Big[\Delta \sum_{k=0}^{K-1} \{ -2 x_{3m}(k) g_m(k) + c_{1m} [\nabla s_m(k)]^2 + B \Delta g_m(k) \sum_{j=1}^{K-1} R(k,j) g_m(j) \} + \Delta \sum_{k=0}^{K-2} c_{2m} [\nabla^2 s_m(k)]^2 \Big] \tag{24}$$

where $x_{3m}(k)$ are the channel output variables. Defining $x_{2m}(k) = \nabla s_m(k)$ and control variables $u_m(k) = \nabla x_{2m}(k)$, $v_m(k) = g_m(k)$, the constraint equations are

$$\begin{aligned} x_{sm}(k+1) &= x_{sm}(k) + \Delta s_m^2(k) \\ x_{gm}(k+1) &= x_{gm}(k) + \Delta v_m^2(k) \\ s_m(k+1) &= s_m(k) + \Delta x_{2m}(k) \\ x_{2m}(k+1) &= x_{2m}(k) + \Delta u_m(k) \end{aligned} \tag{25a}$$

$$x_{3m}(k+1) = a \Delta s_m(k+1) + e^{-\beta \Delta} x_{3m}(k) \tag{25b}$$

where α and β are the channel parameters as in Reference 1.

For the case of white Gaussian noise, the Hamiltonian in terms of Eqs. (24) and (25) is given by

$$H(k) = \sum_{m=1}^{M} \Big\{ \Delta \Big[2 x_{3m} v_m(k) - c_{1m} x_{2m}^2(k) - c_{2m} u_m^2(k) - \frac{BN_0}{2} v_m^2(k) \Big] + \psi \big[x_{3m}(k) + s_m^2(k) \big] + \phi \big[x_{gm}(k) + v_m^2(k) \big] + p_{1m}(k) \big[s_m(k) + \Delta x_{2m}(k) \big] + p_{2m}(k+1) \big[x_{2m}(k) + \Delta u_m(k) \big] + p_{3m}(k+1) \big[\alpha \Delta s_m(k) + \alpha \Delta^2 x_{2m}(k) + e^{-\beta\Delta} x_{3m}(k) \big] \Big\} \tag{26}$$

Maximizing Eq. (26) with respect to $u_m(k)$ and $v_m(k)$ results in

$$v_m(k) = \frac{2}{(BN_0 - 2\phi)} x_{3m}(k) \quad , \quad B \geq 0 \tag{27}$$

$$u_m(k) = \frac{1}{2 c_{2m}} p_m(k+1) \quad , \quad c_{2m} \geq 0 \tag{28}$$

It can be shown that the adjoint equations reduce to M identical sets of coupled linear difference equations with constant coefficients of the form

$$\begin{aligned}
s_m(k+1) &= s_m(k) + \Delta x_{2m}(k) \\
x_{2m}(k+1) &= \frac{\Delta^3 \psi}{c_{2m}} s_m(k) + \Big[1 + \frac{\Delta^2 c_{1m}}{c_{2m}} \Big] x_{2m}(k) - \frac{\Delta^2}{2 c_{2m}} p_{1m}(k) + \frac{\Delta}{2 c_{2m}} p_{2m}(k) \\
x_{3m}(k+1) &= \alpha \Delta s_m(k) + \alpha \Delta^2 x_{2m}(k) + e^{-\beta\Delta} x_{3m}(k) \\
p_{1m}(k+1) &= -2 \Delta \psi s_m(k) - \alpha \Delta W e^{\beta\Delta} x_{3m}(k) + p_{1m}(k) - \alpha \Delta e^{\beta\Delta} p_{3m}(k) \\
p_{3m}(k+1) &= W e^{\beta\Delta} x_{3m}(k) + e^{\beta\Delta} p_{3m}(k)
\end{aligned} \tag{29}$$

where

$$W = 4\Delta [2\phi - BN_0]^{-1} \quad .$$

C. Numerical Examples

As a numerical example of the above formulation consider the solution of Eq. (29) for M = 4 (quaternary system). The parameter values and boundary conditions yielding

mutually orthogonal channel output signals, i.e., those satisfying the relationship

$$C_{ij} = \Delta \sum_{k=0}^{K} x_i(k)\, g_j(k) = c\, \delta_{ij} \quad , \quad i,j = 1,2,\cdots,M \tag{30}$$

where δ_{ij} is the Kronecker delta, are given in Table I. The solutions are illustrated in Figs 2 and 3 for equal energy input signals and a transmission interval T = 2. The last two parameter values are chosen to permit a comparison of system performance with a binary system on the basis of equal data transmission rate. From the figures, it is seen that the direct solution of the problem yields a signalling alphabet with a large spread in frequency. In this form the solution is not useful and a modification is needed.

TABLE I. Parameter values and boundary conditions for Quaternary orthogonal solution of Equations (29).

Solution	$c_1 = c_2$	ψ	$x_2(k)$	Gabor BW	$\rho(K)$
s_1	.01	1.0	1.20	8.34	.440
s_2	.01	1.0	-1.46	20.85	.754
s_3	.01	9.4	1.45	36.05	1.098
s_4	.01	24.0	-1.64	56.09	1.435

K = 30, T = 2, E_s = 10, and all other constants equal to unity.

For an M-ary signal alphabet, the overall Gabor bandwidth may be defined either as

$$B_{max} = \max(B_1, B_2, \cdots, B_M) \tag{31}$$

or

$$B_{rms} = \left[\frac{1}{M} \sum_{m=1}^{M} B_m^2\right]^{1/2} \tag{32}$$

where

$$B_m^2 = \Delta \sum_{k=0}^{K-1} [\nabla s_m(k)]^2 \quad \text{and} \quad m = 1,2,\cdots,M.$$

Nuttall and Amoroso[7] have shown that alphabets of equal bandwidth signals which minimize both Eqs. (31) and (32) can be obtained from a set of harmonic signals by means of orthogonal transformations proportional to the Hadamard matrices. It is known that Hadamard matrices exist for M = 2, 4, 8, 12, $\cdots$, 200 with few exceptions (see Ref. 9, theorem 4.4). For M = 4, there are two nonequivalent forms of the orthogonal matrix

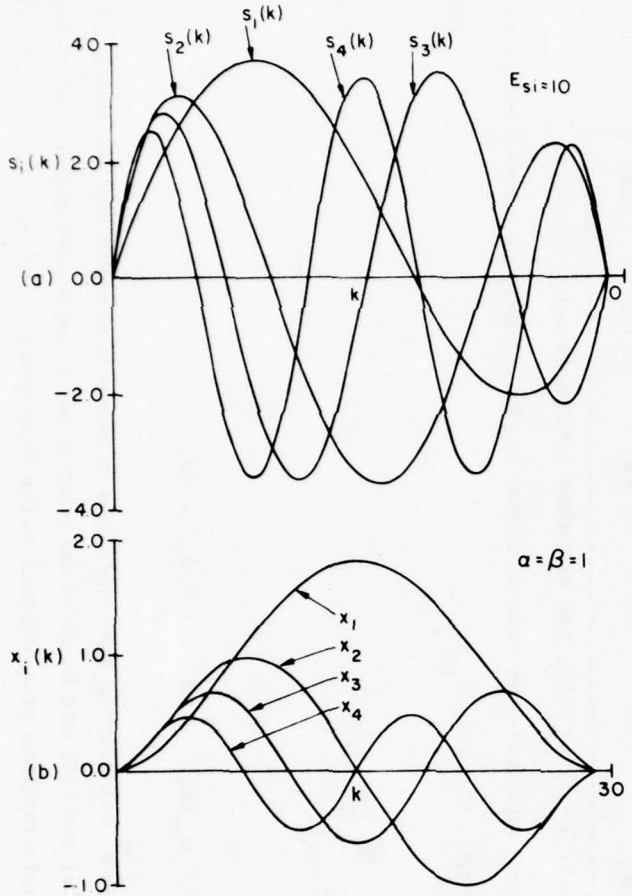


Fig. 2. Envelopes of discrete signals and corresponding discrete channel outputs for quaternary orthogonal signalling with K = 30.

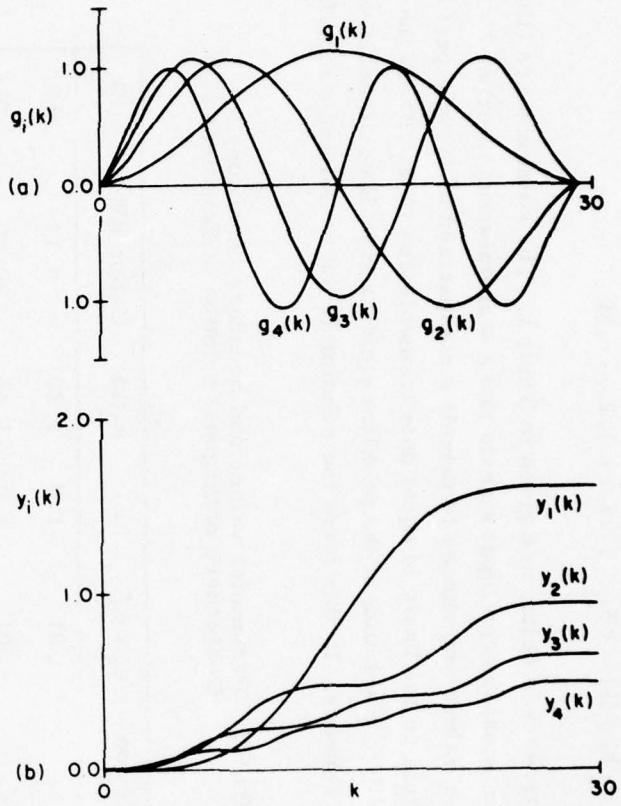


Fig. 3. Envelopes of discrete correlation functions and corresponding detector outputs for quaternary orthogonal signalling with K = 30.

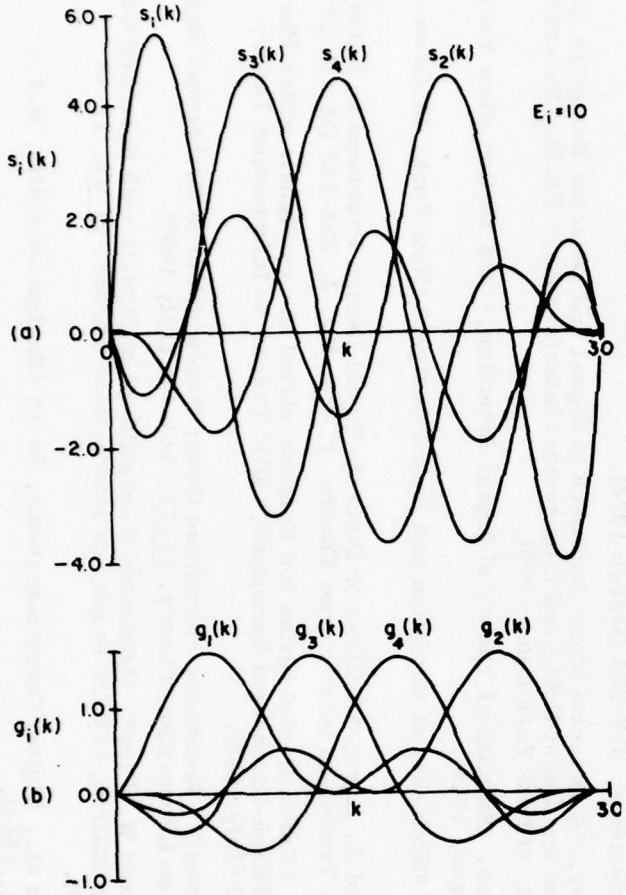


Fig. 4. Systems of Figs. 2 and 3 transformed via the Hadamard orthogonal transformation Equation (33).

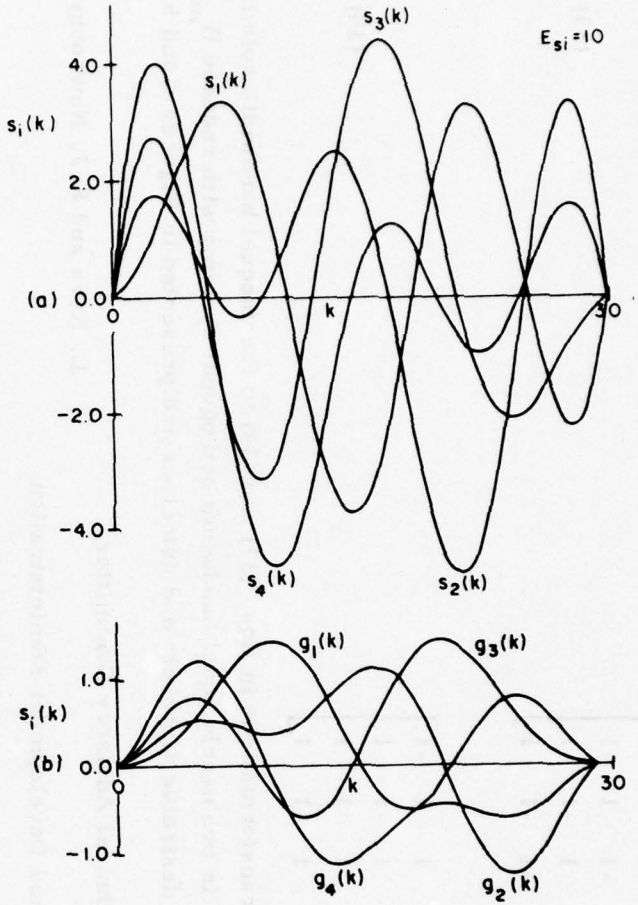


Fig. 5. Systems of Figs. 2 and 3 transformed via the Hadamard orthogonal transformation Equation (34).

given by

$$U_1 = \frac{1}{\sqrt{4}} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \tag{33}$$

and

$$U_2 = \frac{1}{\sqrt{4}} \begin{bmatrix} 1 & 1 & 1 & -1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 \\ -1 & 1 & 1 & 1 \end{bmatrix} \tag{34}$$

An application of transformations in Eqs. (33) and (34) to the unequal bandwidth solution of Eq. (29) results in two nearly equal bandwidth orthogonal systems with reduced B_{max} and B_{rms}. These desirable alphabets and detectors are presented in Figures 4 and 5.

Joint Services Technical Advisory Committee
F44620-74-C-0056

L. Kurz and R.J. Nawrocky

Energy Research and Development Administration

REFERENCES

1. R.J. Nawrocky and L. Kurz, *"Design of Digital Filters for Extraction of Signals* from Intersymbol Interference and Noise," Proc. 8th Annual Conf. on Information Sciences and Systems, 499-503 (March 1974).

2. R.J. Nawrocky, "A Discrete Time Approach to Signal and Detector Design in Additive Gaussian and Burst Noise and Intersymbol Interference," Ph.D. Dissertation, Polytech. Inst. of New York (June 1975).

3. C.W. Helstrom, "Statistical Theory of Signal Detection," 2nd Edition (New York: Pergamon Press, 1968).

4. C.L. Weber, "Elements of Detection and Signal Design," (New York: McGraw-Hill, 1968).

5. H.I. Silver and L. Kurz, "A Class of Discrete Signal Design Problems in Burst Noise," IEEE Trans. on Information Theory, IT-18, No. 2, 258-262 (March 1972).

6. A.H. Nuttall, "Error Probabilities for Equicorrelated M-ary Signals under Phase-coherent and Phase-incoherent Reception," IRE Trans. on Information Theory, IT-8, 305-314 (July 1962).

7. A.H. Nuttall and F. Amaroso, "Minimum Gabor Bandwidth of M Orthogonal Signals," IEEE Trans. on Information Theory, IT-11, 440-445 (July 1965).

8. A.H. Nuttall and W. Floyd, "Minimum Bandwidth of M Signals with Specified Code or Correlation Matrix," IEEE Trans. on Information Theory, IT-15, 38-48 (January 1969).

9. S. Golomb, et al., Digital Communications, 54-55 (Englewood Cliffs, N.J.: Prentice-Hall, 1964).

A GENERALIZED QUANTILE DETECTOR

J.I. Cochrane and L. Kurz

In this report, a particularly robust subset of quantile detectors is considered. Unlike the rank detectors, the detectors considered here are easy to implement. The knowledge of quantile required for proper operation of the detector can be obtained using one of the methods suggested by Kersten and Kurz.[1,2]

A. Mathematical Formulation and Asymptotic Properties

Consider n_i i.i.d. samples from the distributions F_i, $i = 1, 2, \cdots, M$. Form the empirical distribution functions

$$F_{in_i}(x) = n_i^{-1} \sum_{j=1}^{n_i} u(x - x_{ji})$$

where $u(\cdot)$ is the unit step function and x_{ji} is the j-th sample from F_i. Define the weighted sum of empirical distribution functions

$$H_N(x) = \sum_{i=1}^{M} \beta_i F_{in_i}(x)$$

where

$$\sum_{i=1}^{M} \beta_i = 1, \quad \beta_i \geq 0, \quad \sum_{i=1}^{M} n_i = N$$

Form a class of tests

$$T_i = \sum_{j=1}^{m} C_j W_i(\lambda_j) \tag{1}$$

where

$$0 < \lambda_1 < \cdots < \lambda_m < 1$$

$$W_i(\lambda_j) = F_{in_i} H_N^{-1}(\lambda_j)$$

and C_j are appropriately selected constants. T_i of Eq. (1) is a general representation for a broad class of threshold tests. Depending on the choice of β_i and the relative growth rates of the n_i, Eq. (1) may be used for asymptotic analysis of quantile and joint quantile tests with random and fixed partitions. From the partition viewpoint, the number of thresholds m is a measure of the robustness of the test. Since for small m the tests have still high efficacy, the advantages of simplicity and robustness of the

tests given by Eq. (1) is even further accentuated. In particular for $\beta_i = 0$, T_i can be formed by a simple threshold process after the computation of H_N, and no ranking of x_{ji} is required.

Theorem: The set $W_i(\lambda_j)$, $j = 1, \cdots, m$, converges in law as $n_{min} \to \infty$ to a set of jointly normal random variables $W_{io}(\lambda_j)$, where

$$E\left[W_{io}(\lambda_j)\right] = F_i H^{-1}(\lambda_j) \tag{2}$$

and

$$\operatorname{cov}\left\{W_{io}(\lambda_j),\ W_{io}(\lambda_k)\right\} = \sum_{\mu=1}^{M} n_\mu^{-1}(\delta_{i_\mu} - \beta_\mu a_i(\lambda_j)) \cdot (\delta_{i_\mu} - \beta_\mu a_i(\lambda_k)) F_\mu H^{-1}(\lambda_j)\left[1 - F_\mu H^{-1}(\lambda_k)\right] \tag{3}$$

for $\lambda_i < \lambda_j$.

Proof: It follows directly from Lemma 1 of the Appendix in Reference 3.

It can be shown that the efficacy of the T_i tests is of the form

$$\mathcal{E}(T_i) = \frac{\underset{\sim}{C}^t B \underset{\sim}{C}}{\underset{\sim}{C}^t \sum_o \underset{\sim}{C}} \tag{4}$$

where $\underset{\sim}{C}^t = [C_1, C_2, \cdots, C_m]$ and B, $\sum_o$ are m x m symmetric matrices with elements

$$b_{jk} = \left.\frac{\partial W_i(\lambda_j)}{\partial \theta}\right|_{\theta=0} \cdot \left.\frac{\partial W_i(\lambda_k)}{\partial \theta}\right|_{\theta=0} \tag{5}$$

$$\sigma_{jk} = \operatorname{cor}\left\{W_i(\lambda_j),\ W_i(\lambda_k)\right\} \tag{6}$$

The scores which maximize Eq. (4) and the corresponding efficacy may be calculated by forming a linear transformation

$$\mathcal{E}(T_i) = \underset{\sim}{Y}_N^t \left(\sum^{-1/2} B \sum^{-1/2}\right) \underset{\sim}{Y}_N \tag{7}$$

where

$$\underset{\sim}{Y}_N = \sum^{-1/2} \underset{\sim}{C}_N \qquad \underset{\sim}{Y}^t \underset{\sim}{Y} = 1$$

The maximum efficacy equals the largest eigenvalue of $B \sum^{-1}$ and the optimum scores of the test are the corresponding eigenvector. Since B is of rank 1, Eq. (7) reduces to

$$\mathcal{E}(T_i) = \underset{\sim}{F}^t \sum^{-1} \underset{\sim}{F} \tag{8}$$

where

$$\underset{\sim}{F}^t = \left[f_1, \cdots, f_m\right]$$

and

$$f_k = \frac{\partial}{\partial\theta} F_i H^{-1}(\lambda_k)\Big|_{\theta=0} \tag{9}$$

B. Numerical Results

To illustrate the performance of the T_i statistics, efficacy calculations have been carried out for the Lehmann alternative, the Gaussian scale and shift alternatives, and the Cauchy scale and shift alternatives. The Cauchy alternatives are chosen to illustrate the performance of the classifiers in an extreme case where the moments of the distribution do not exist. The properties of the Lehmann alternative and its relation to the shift alternative is discussed elsewhere in this report.

In Table I, the efficacy of the T_i statistic for M = 2 is shown for increasing values of m where the parameter $(n_1+n_2)/n_1n_2$ has been suppressed. In each case the set of generalized quantiles $\lambda_1, \cdots, \lambda_m$ has been selected to maximize the efficacy. The last column is the efficacy of the locally most powerful test for the particular case.

TABLE I. Maximum efficacy of T_i statistics.

Alt. \ M	1	2	3	4	5	6	ε_{max}
Gaussian Shift	.64	.810	.883	.920	.942	.956	1.
Cauchy Shift	.405	.429	.434	.470	.479	.483	.5
Lehmann Alt.	.646	.820	.891	.927	.948	.961	1.
Gaussian Scale	.61	1.30	--	1.64	--	1.76	2.
Cauchy Scale	.143	.405	--	.461	--	.479	1.5

The efficacy of the best T_i test reaches 94% of the locally most powerful test for $m = 5$ for the shift and Lehmann alternatives, and 82% for the Gaussian scale alternatives for $m = 4$.

The efficacy of the T_i tests for $M = 2$ is independent of parameters β_1 and β_2. By letting $\beta_1 = 0$ we see that the two-sample quantile test has the same efficacy as the rank test while offering increased simplicity.

Tables II and III show the optimum $\underset{\sim}{\lambda}$ and $\underset{\sim}{C}$ for the examples calculated. The scores have been normalized by $C^tC = 1$. Table IV is included to demonstrate the robustness of the T_i tests. In this table the test designed for one hypothesis testing problem (detection problem) has been applied to a different noise environment. The resulting loss compared to the best T_i test for that environment is tabulated.

C. Finite Sample Size Analysis

In this section, the exact moments of $W_i(\lambda) = F_{in_i}(H_n^{-1}(\lambda))$ under the Lehmann alternative ($M = 2$) are expressed in computationally convenient form. Since the T_i statistics are linear functionals of $W_i(\lambda)$, these computations are useful in comparing finite sample size and asymptotic behavior of the tests.

The moments of $W_1(\lambda)$ are calculated by noting that it can be expressed as a sum of binary random variables

$$W_1(\lambda) = n_1^{-1} \sum_{j=1}^{n_1} u\left(x_{(j)} - y_{(f(j))}\right) \tag{10}$$

where $x_{(j)}$ is the j-th order statistic from F_1 and $y_{(j)}$ is the j-th order statistic from F_2 and $f(\cdot)$ is a nondecreasing integer function of j. Since $f(\cdot)$ is a nonincreasing function, the joint probabilities which appear in the calculation of the higher moments simplify to allow exact computation. If F_1 and F_2 are absolutely continuous, it can be shown that the moments are

$$E\left[W_1^m(\lambda)\right] = \sum_{k=1}^{n_1} A_m(k) P_{\lambda,\beta}(k) \tag{11}$$

where

$$A_m(k) = n_1^{-m}\left[k^m - (k-1)^m\right] \tag{12}$$

and

TABLE II. Optimum threshold settings for T_i statistics.

Alt.	M	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
Gaussian Shift	1	.5	--	--	--	--	--
	2	.27	.73	--	--	--	--
	3	.165	.5	.835	--	--	--
	4	.108	.350	.650	.892	--	--
	5	.075	.255	.500	.745	.925	--
	6	.055	.190	.390	.925	.81	.945
Cauchy Shift	1	.5	--	--	--	--	--
	2	.445	.555	--	--	--	--
	3	.080	.500	.920	--	--	--
	4	.090	.435	.565	.91	--	--
	5	.098	.403	.500	.597	.902	--
	6	.100	.385	.465	.535	.615	.900
Lehmann Alt.	1	.205	--	--	--	--	--
	2	.073	.360	--	--	--	--
	3	.035	.170	.470	--	--	--
	4	.019	.093	.258	.548	--	--
	5	.011	.055	.153	.328	.603	--
	6	.008	.038	.103	.220	.400	.655
Gaussian Scale	1	.95 or .05	--	--	--	--	--
	2	.07	.93	--	--	--	--
	4	.02	.14	.86	.98	--	--
	6	.011	.06	.17	.83	.94	.989
Cauchy Scale	1	.2 or .8	--	--	--	--	--
	2	.25	.75	--	--	--	--
	4	.186	.314	.685	.814	--	--
	6	.153	.250	.348	.652	.750	.847

TABLE III. Scores which maximize efficacy with optimum threshold settings.

Alt.	M	C_1	C_2	C_3	C_4	C_5	C_6
Gaussian Shift	2	.707	.707	--	--	--	--
	3	.605	.518	.605	--	--	--
	4	.553	.419	.419	.553	--	--
	5	.524	.400	.373	.400	.524	--
	6	.501	.371	.333	.333	.371	.501
Cauchy Shift	2	.707	.707	--	--	--	--
	3	-.300	.910	-.300	--	--	--
	4	-.390	.590	.590	-.390	--	--
	5	-.441	.443	.469	.443	-.441	--
	6	-.479	.356	.380	.380	.356	-.479
Lehmann Alt.	2	.857	.510	--	--	--	--
	3	.808	.480	.344	--	--	--
	4	.780	.464	.332	.258	--	--
	5	.763	.453	.324	.253	.208	--
	6	.752	.444	.321	.252	.204	--
Gaussian Scale	2	-.707	.707	--	--	--	--
	4	-.636	-.309	.309	.636	--	--
	6	-.597	-.317	-.206	.206	.317	.597
Cauchy Scale	2	-.707	.707	--	--	--	--
	4	-.5	-.5	.5	.5	--	--
	6	-.4	-.43	-.4	.4	.43	.4

TABLE IV. Loss in dB for three environments.

	Optimum Settings for Lehmann Alternatives		Optimum Settings for Gaussian Shift Alternative		Optimum Settings for Cauchy Shift Alternative	
M	Gaussian Noise	Cauchy Noise	Lehmann Alt.	Cauchy Noise	Lehmann Alt.	Gaussian Noise
1	1.11	5.65	1.29	0	1.29	0
2	1.04	4.74	1.15	2.55	1.89	0.62
3	1.00	4.45	1.08	2.74	3.36	3.20
4	1.06	4.76	1.02	3.24	4.47	2.99
5	0.96	4.89	0.99	3.36	4.47	2.99
6	0.93	4.96	0.97	3.44	4.47	3.01

$$P_{\lambda,\beta}(k) = n_2\binom{n_2-1}{f(k)-1}\sum_{j=k}^{n_1}(-1)^{j-k}\binom{n_1}{j}\binom{j-1}{k-1}B\left((1+\varepsilon)j + f(k),\ n_2 - f(k) + 1\right) \qquad (13)$$

if $1 \le f(k) \le n_2$, $F_2 = F_1^{1+\varepsilon}$

Otherwise

$$P_{\lambda,\beta}(k) = 1 \quad \text{if} \quad f(k) > n_2$$

$$= 0 \quad \text{if} \quad f(k) < 1$$

$B(\cdot,\cdot)$ is the tabulated Beta function and $f(k)$ is the smallest integer greater than or equal to

$$\frac{n_2}{1-\beta_1}\left(\lambda - \frac{\beta_1(k-1)}{n_1}\right) \qquad (14)$$

Joint Services Technical Advisory Committee
F44620-74-C-0056

J.I. Cochrane and L. Kurz

Bell Telephone Laboratories
Doctoral Support Plan

REFERENCES

1. P. Kersten and L. Kurz, "Robustized Vector Robbins-Monro Algorithm with Applications to M-Interval Detection," Information Sciences, to appear.

2. P. Kersten and L. Kurz, "Bivariate m-Interval Classifiers with Application to Edge Detection," Information and Control, to appear.

3. L. Kurz, "Nonparametric Detectors Based on Partition Tests," Supplementary Notes, EE704, Polytech. Inst. of New York (September 1975).

THE NYQUIST PROBLEM IN THE PRESENCE OF NONLINEARITIES IN THE DATA TRANSMISSION SYSTEM

L. Kurz and G. Soloway

It has been previously established that the introduction of a memoryless nonlinearity between the channel and the receiver filter, for the purpose of suppressing impulsive noise, is going to complicate the problem of intersymbol interference.[1] It is no longer possible to achieve perfect equalization by the application of Nyquist's criterion, although for most reasonable signals, the intersymbol interference will be bounded.

In this report, a frequency domain approach to equalization in data transmission system in the presence of memoryless nonlinearity is considered, with particular stress on elimination of intersymbol interference caused by the most troublesome terms.

A. A Frequency Domain Receiver Design Procedure

Results of Ref. 1 suggest that the linear receiving filter (equalizer) can be designed for an input signal which is a sum of the linear and cubic terms. This has one major drawback. The relative amplitudes of the linear and cubic terms are not invariant. They are a function of the nonlinearity, signal amplitude and noise parameters. It is highly undesirable to use a design procedure which would change as the variance of the noise varies.

Theorem 1 of Ref. 1 suggests a method by which this problem may be alleviated. Let the desired part of the signal, usually the linear part, have a spectrum $G_A(\omega)$. Let the undesired part of the spectrum, perhaps the cubic term, have a spectrum $G_B(\omega)$. From Nyquist's criterion and Theorem 1 of Ref. 1, the following constraint equations can be written

$$\sum_k G_A^{(k)}(\omega)\, G_R^{(k)}(\omega) = Tx_o \qquad |\omega| \le \frac{\pi}{T} \tag{1}$$

$$\sum_k G_B^{(A)}(\omega)\, G_R^{(k)}(\omega) = 0 \qquad |\omega| \le \frac{\pi}{2} \tag{2}$$

where the notation of Ref. 2 has been used. One can then seek the minimum of

$$\sigma^2 = \frac{N_o}{2\pi} \int_{-\pi/T}^{\pi/T} \sum_k |G_k^{(k)}(\omega)|^2 \, d\omega \tag{3}$$

Thus, for each value of ω, $|\omega| \le \frac{\pi}{T}$, one seeks the stationary point of

$$J = \sum_k \left[|G_R^{(k)}(\omega)|^2 + \lambda(\omega) G_A^{(k)}(\omega) G_R^{(k)}(\omega) + \nu(\omega) G_B^{(k)}(\omega) G_R^{(k)}(\omega) \right]$$

from which one arrives at

$$G_R^{(k)}(\omega) = \frac{-\lambda(\omega)}{2}\left[G_A^{(k)}(\omega)\right]^* - \frac{\nu(\omega)}{2}\left[G_B^{(k)}(\omega)\right]^* \tag{4}$$

where * denotes complex conjugate.

Using Eq. (4) in Eqs. (1) and (2) yields

$$\sum_k\left[\frac{-\lambda(\omega)}{2}\left|G_A^{(k)}(\omega)\right|^2 - \frac{\nu(\omega)}{2}G_A^{(k)}(\omega)\left[G_B^{(k)}(\omega)\right]^*\right] = Tx_o \tag{5}$$

$$\sum_k\left\{\frac{-\lambda(\omega)}{2}\left[G_A^{(k)}(\omega)\right]^* G_B^{(k)}(\omega) - \frac{\nu(\omega)}{2}\left|G_B^{(k)}(\omega)\right|^2\right\} = 0 \tag{6}$$

Solving Eq. (6) for $\nu(\omega)$

$$\nu(\omega) = \frac{-\sum_k \lambda(\omega)\left[G_A^{(k)}(\omega)\right]^* G_B^{(k)}(\omega)}{\sum_k\left|G_B^{(k)}(\omega)\right|^2} \tag{7}$$

Using Eq. (7) in Eq. (5) and solving for $\lambda(\omega)$ yields

$$\lambda(\omega) = \frac{Tx_o}{\sum_k\left\{-\frac{1}{2}\left|G_A^{(k)}(\omega)\right|^2 + \left[\frac{\sum_i G_A^{(i)}(\omega)\,G_B^{(i)}(\omega)}{2\sum_i\left|G_B^{(i)}(\omega)\right|^2}\right]G_A^{(k)}(\omega)\left[G_B^{(k)}(\omega)\right]^*\right\}} \tag{9}$$

Similarly, solving for $\lambda(\omega)$ from Eq. (6) and using this in Eq. (5) yields the following expression for $\nu(\omega)$

$$\nu(\omega) = \frac{Tx_o}{\sum_k\left\{\left[\frac{\sum_i\left|G_B^{(i)}(\omega)\right|^2}{2\sum_i\left[G_A^{(i)}(\omega)\right]^* G_B^{(i)}(\omega)}\right]\left|G_A^{(k)}(\omega)\right|^2 - \frac{1}{2}G_A^{(k)}\left[G_B^{(k)}(\omega)\right]^*\right\}} \tag{10}$$

Substituting from Eqs. (9) and (10) in Eq. (4) results in an expression for the desired filter

$$G_R^{(k)} = \frac{Tx_o [G_A^{(k)}(\omega)]^*}{\sum_k \left\{ |G_A^{(k)}(\omega)|^2 - \left[\frac{\sum_i [G_A^{(i)}(\omega)]^* G_B^{(i)}(\omega)}{\sum_i |G_B^{(i)}(\omega)|^2} \right] G_A^{(k)}(\omega) [G_B^{(k)}(\omega)]^* \right\}}$$

$$+ \frac{Tx_o [G_B^{(k)}(\omega)]^*}{\sum_k \left\{ G_A^{(k)}(\omega) [G_B^{(k)}(\omega)]^* - \left[\frac{\sum_i |G_B^{(i)}(\omega)|^2}{\sum_i [G_A^{(i)}(\omega)]^* G_B^{(i)}(\omega)} \right] |G_A^{(k)}(\omega)|^2 \right\}} \qquad (11)$$

It is possible to extend this result to include more terms. The number of terms for which intersymbol interference can be eliminated depends upon the bandwidth available. The cubic device produces power terms and cross-product terms with three times the original signal bandwidth. This extra bandwidth introduces additional degrees of freedom in the receiver design, allowing for additional constraints to be satisfied.

B. Examples

An example will help clarify the use of the filter design procedure. As in standard PAM systems, the signal is assumed to have a frequency response corresponding to the square root of a raised cosine. Here the excess bandwidth was taken to be .5. This frequency response and that of its corresponding time response cubed are shown in Figure 1. Equation (11) may now be solved for the desired filter frequency response. This filter characteristic is shown in Figure 1. The time responses which result from passing the linear and cubic terms through this filter are shown in Figure 2.

It may easily be seen that the design satisfies all the constraints. However, this result has not been achieved without penalty. The time responses are not well damped. This will increase jitter problems. Also, a glance at the filter response reveals that this filter permits far greater amount of noise to pass through. Therefore, the system designer must make a choice whether it is more important to eliminate the distortion due to the cubic term, or more effectively suppress the noise.

An alternative would be to control the intersymbol interference due to the cubic term while not reducing the intrasymbol interference to zero. To accomplish this the right hand side of constraining Eq. (2) would be set equal to a constant, say Tx_o'. Following the same procedure which resulted in Eq. (11), one obtains

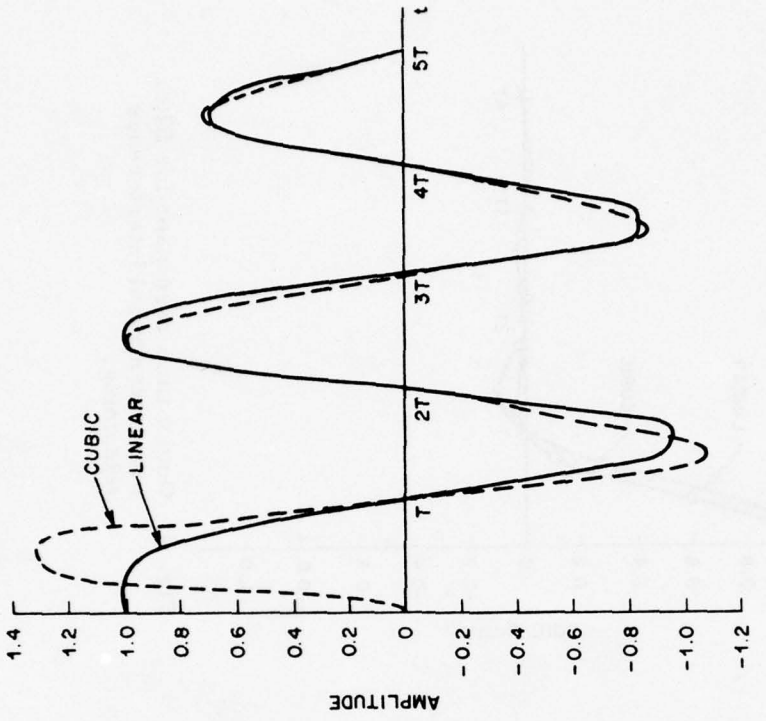


Fig. 2. Output time responses due to linear and cubic terms.

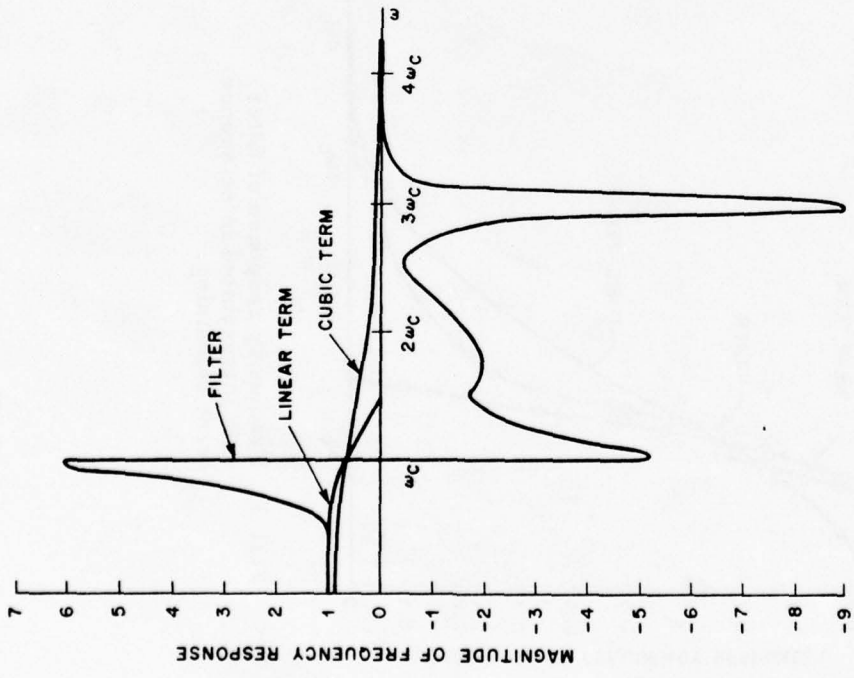


Fig. 1. Frequency responses of signals and filter.

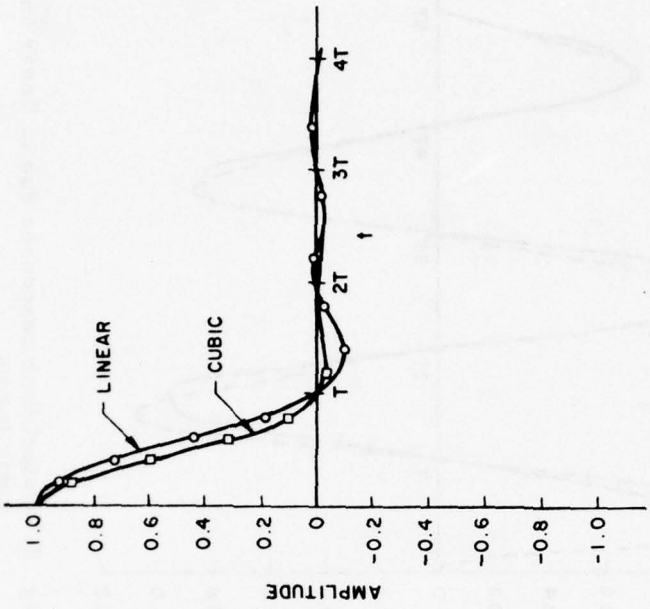


Fig. 4. Output time responses for filter with intrasymbol interference remaining.

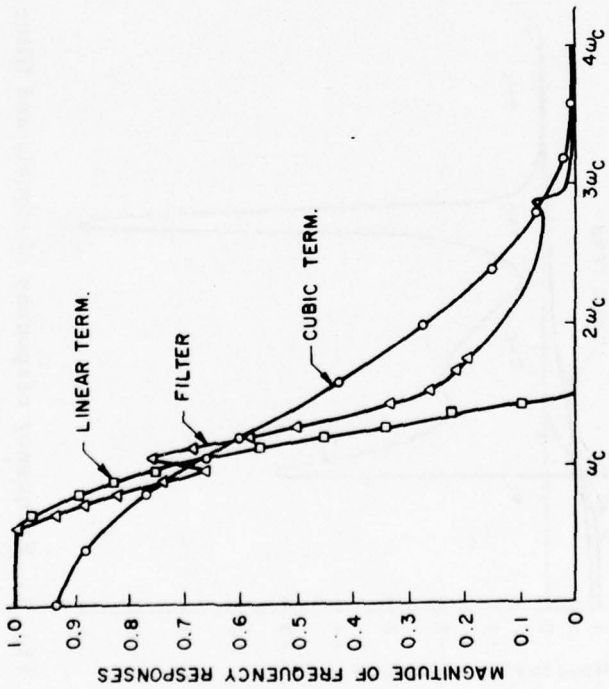


Fig. 3. Frequency response of filter with intrasymbol interference term remaining.

$$G_R^{(k)}(\omega) = \frac{\left\{ Tx_o - Tx'_o \dfrac{\sum_i G_A^{(i)}(\omega)\,[G_B^{(i)}(\omega)]^*}{\sum_i |G_B^{(i)}(\omega)|^2} \right\} [G_A^{(k)}(\omega)]^*}{\sum_k \left\{ |G_A^{(k)}(\omega)|^2 - \left[\dfrac{\sum_i [G_A^{(i)}(\omega)]^* G_B^{(i)}(\omega)}{\sum_i |G_B^{(i)}(\omega)|^2} \right] G_A^{(k)}(\omega)\,[G_B^{(k)}(\omega)]^* \right\}}$$

$$+ \frac{\left\{ Tx_o - Tx'_o \dfrac{\sum_i |G_A^{(i)}(\omega)|^2}{\sum_i [G_A^{(i)}(\omega)]^* G_B^{(i)}(\omega)} \right\} [G_B^{(k)}(\omega)]^*}{\sum_k \left\{ |G_A^{(k)}(\omega)|^2 [G_B^{(k)}(\omega)]^* - \left[\dfrac{\sum_i |G_B^{(i)}(\omega)|^2}{\sum_i [G_A^{(i)}(\omega)]^2 G_B^{(i)}(\omega)} \right] |G_A^{(k)}(\omega)|^2 \right\}} \qquad (12)$$

The design problem considered in the first part of this section is repeated using Equation (12). The filter frequency response is shown in Figure 3. It is seen that this represents a far more practical solution. This filter will have superior noise performance compared to that of Figure 1. Furthermore, the time responses are well damped (see Figure 4). Its only drawback is that a small intrasymbol interference term remains which is a function of the nonlinearity, the noise, and the signal amplitude.

Joint Services Technical Advisory Committee
F44620-74-C-0056

L. Kurz and G. Soloway

Bell Telephone Laboratories

Doctoral Support Plan

REFERENCES

1. L. Kurz and G. Soloway, "The Effect of Nonlinearities on Suppression of Intersymbol Interference," Progress Report No. 40 to JSTAC, Polytech. Inst. of New York, Report No. R-452.40-75, 402-406 (November 1975).

2. R.W. Lucky, J. Salz and E.J. Weldon, Jr., "Principles of Data Communications," (New York: McGraw-Hill, 1968).

EDGE DETECTION USING ANOVA TECHNIQUES IN CONJUNCTION WITH QUANTILE STATISTICS

L. Kurz and P. Legakis

In this report, analysis of variance (ANOVA) techniques are employed for the solution of the edge detection problem. Although ANOVA techniques have been applied to edge detection problems before,[1,5] the procedures presented here take full advantage of the data reduction based on estimated quantiles. Specifically, techniques based on the one-way analysis of variance based on quantile estimates are used. The quantile estimation procedures yield asymptotically normal estimates regardless of the distribution of the observables. In addition, by properly selecting the quantiles, the much simpler fixed effects model of analysis of variance can be applied to gray level and texture edge detection.

A. Statement of the Analysis of Variance Problem

Let Q be an $L_m \times L_n$ array of quantiles resulting from an array H of image intensities after some form of quantile estimation on H. Let q_{rs} be the quantile estimate based on the sample $\eta_{r,s-n/2} \cdots \eta_{r,s+n/2}$ of size n, taken from the array of image intensities H. Assuming a rectangular scan on the original data, the order that the quantiles are transmitted by the estimator is

$$q_{11}, q_{12}, \cdots, q_{1L_n}, q_{21}, \cdots, q_{2L_n}, \cdots, q_{L_{mi}}, \cdots, q_{L_m L_n} \tag{1}$$

Each of the elements q_{rs} may be thought of as a random variable with mean μ_{rs} and variance σ^2_{rs}. Furthermore, the random variables q_{rs} are asymptotically normally distributed regardless of whether stochastic approximation[2,3] or the order statistic[4] approach is used in their estimation. The means have different meaning depending on what the quantiles are. For example, if $\{q_{rs}\}$ estimate the 50th percentile, is the mean due to the presence of gray level at the point (r, s) and σ^2_{rs} is the variance which accounts for zero mean additive noise due to finite-scanner bandwidth, image quantization, transmission, and image texture. If q_{rs} is a high quantile, e.g., the 95th percentile, the μ_{rs} is the average change in texture of the point (r, s) and the variance has the same significance as for the case where q_{rs} is the 50th percentile estimate. Thus, edges are defined in terms of the values of q_{rs} at either side of the edge. For gray level edges, the quantile estimates are medians and the mean of the medians changes across the two adjacent gray levels while the variance stays the same. For texture edges, the quantile estimates are high percentiles and the mean of the percentiles changes when actual changes in texture are encountered. In either case, one is looking for a shift in the mean of either the 50th or the high (low) percentiles.

B. Edge Detection Using One-Way ANOVA

For each point (r, s) of the $L_m \times L_n$ quantile array Q, a neighborhood is defined consisting of a horizontal array of M groups, each group consisting of N points. The whole neighborhood is centered around the point (r, s), where $r = 1, 2, 3, \cdots, L_m$ and $s = k, k+1, \cdots, L_n - k$. ($k = (M \times N) : 2$ = Largest integer greater or equal to $(M \times N) : 2$) Expressing by ${}_{rs}X$ the array representing the quantile neighborhood of (r, s), then ${}_{rs}X$ consists of elements $\{{}_{rs}X_{ij}\}$ which can be written in terms of the quantile elements as follows:

$$
\begin{aligned}
&{}_{rs}X_{ij} = q_{r;\, s-k+N(i-1)+j} \\
&i = 1, 2, \cdots, \mu;\ j = 1, 2, \cdots, N \\
&r = 1, 2, \cdots, L_m;\ s = k, k+1, \cdots, L_n - k
\end{aligned}
\tag{2}
$$

If for any integer N the notation

$$
\sim N = 1, 2, 3, \cdots, N \tag{3}
$$

is adopted, then a computationally simpler representation of ${}_{rs}X$ results, namely,

$$
\begin{aligned}
&{}_{rs}X = q_{r;\, s-k+a(M \times N)} \\
&r = 1, 2, \cdots, L_m;\ s = k, k+1, \cdots, L_n - k
\end{aligned}
\tag{4}
$$

Note that the expression for the quantiles has one row index and $M \times N$ column indices as required for the formulation of ${}_{rs}X$. Also, the reason why s starts from k is that one wants to form neighborhoods that are centered around the points of interest (r, s) and thus, by necessity, the first and last $(k - 1)$ columns of the quantile array Q are omitted so far as the classification of those particular points is concerned. They are, however, taken into account beginning with column k if the neighborhood includes them. The exclusion of the first and last $(k - 1)$ columns does not constitute any serious drawback for the procedure since the area of interest usually occupies the center of the picture and the scanner resolution is such that k is much smaller than the distance between successive edges.

The array ${}_{rs}X$ can be thought of as a mask that is moved over points (r, s) for a total of $L_m \times (L_n - 2k)$ times for the complete coverage of the quantile array Q. The masking process together with the application of one-way ANOVA techniques will result in $L_m \times (L_n - 2k)$ statistics which give an edge enhanced representation of the original picture.

If an n:1 data reduction ratio is assumed from the array of observations to the array of quantiles, then for fixed image size and variable scanner resolution (N_r, N_s)

on the original data, the minimum distance between successive edges must be a large multiple of $M \times N \times n$ picture elements in order to avoid multiple false edge detections and significant performance degradation. This requirement also helps in the picture left- and right-margin problem because if $N_s >> M \times N \times n > (M \times N \times n) \div 2$, the border points become comparatively insignificant. Notice that one-way ANOVA formulation does not impose any requirements on N_r, the vertical scanner resolution, because the operation proceeds along rows rather than columns. For the formation of neighborhoods that have vertical and diagonal, in addition to horizontal symmetry, one can still use one-way ANOVA with two treatments provided that a reference sample is available. The first treatment consists of the points of the reference sample which is of known classification and the second treatment consists of all the neighborhood points of (r, s) as a group. In the latter case, where a reference sample is available, one is actually performing pattern as opposed to edge detection, and this procedure will give much better results than direct either one- or two-way ANOVA. This should come as no surprise for two reasons. First, the choice of the neighborhood is more representative of the environment of (r, s) and, second, the availability of a reference sample yields additional knowledge that any good detector should take advantage of. Two-dimensional neighborhoods, in addition to the requirements on the scanner resolution N_r in the horizontal or x-direction, impose constraints on the scanner resolution N_s along the vertical or y-direction, $N_r >> M1$, where M1 is the number of rows that are used in the formation of the neighborhood of (r, s). Notice that the data reduction factor n does not impose any constraints on the vertical reduction and this is because quantile estimation takes place along the x-direction only. Large scanner resolutions are desirable because they provide much more information about the scanned image, but they must be moderate enough to make the storage problem and the I/O activity realistic. This excludes the consideration of the processing time of the detection procedure.

So far, scanner resolution requirements and the general form of ${}_{rs}X$ have been discussed, yet the problem of selection of the number of groups of observations M and the number of observations N within each group has been excluded. As long as the neighborhood parameters M and N are greater than unity, ${}_{rs}X$ can be used as a possible array in the one-way ANOVA scheme. Increasing the product of these parameters, decreases the variance of ${}_{rs}f$ and has a beneficial effect on the significance and power. However, if the product is made too large, one is running the risk of multiple detections. The best policy is to make the product $M \times N$ as large as possible subject to the requirement that it should also be much smaller than the distance between successive edges. Since no feasible analytical procedure for selecting optimum values of M and N has been found, the problem is easily resolved via simulation.

In the application of the one-way ANOVA techniques for edge detection, arrays ${}_{rs}X$ are formed as indicated above, and the problem is posed as in a hypothesis testing framework as follows:

H: groups are homogeneous → No edge

K: groups are heterogeneous → Edge

The pertinent statistics for the point (r, s) can be obtained through iteration by varying r and s over their respective ranges.

$$ {}_{rs}\bar{x}_{i.} = \frac{1}{N} \sum_{j=1}^{N} q_{r;\, s-k+N(i-1)+j} \tag{5} $$

$$ {}_{rs}\bar{x}_{..} = \frac{1}{M} \sum_{i=1}^{M} {}_{rs}\bar{x}_{i.} = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} q_{r;\, s-k+N(i-1)+j} \tag{6} $$

$$ {}_{rs}S_a^2 = N \sum_{i=1}^{M} \sum_{j=1}^{N} ({}_{rs}\bar{x}_{i.} - {}_{rs}\bar{x}_{..})^2 \tag{7} $$

$$ {}_{rs}S_e^2 = \sum_{i=1}^{M} \sum_{j=1}^{N} (q_{r;\, s-k+N(i-1)+j} - {}_{rs}\bar{x}_{i.})^2 \tag{8} $$

$$ {}_{rs}S_T^2 = {}_{rs}S_a^2 + {}_{rs}S_e^2 \tag{9} $$

$$ {}_{rs}f = \frac{\dfrac{{}_{rs}S_a^2}{(M-1)}}{\dfrac{{}_{rs}S_e^2}{M(N-1)}} = \frac{M(N-1)}{(M-1)} \frac{{}_{rs}S_a^2}{{}_{rs}S_e^2} \tag{10} $$

Since the quantiles are asymptotically normally distributed

$$ \text{H: } {}_{rs}f \sim {}_{rs}F(M-1),\, M(N-1) \tag{11} $$

$$ \text{K: } {}_{rs}f \sim {}_{rs}F(M-1),\, M(N-1)\,(8^2) \tag{12} $$

where ${}_{rs}F(M-1),\, M(N-1)$ again denotes a central F and $F_{[\ ]}(8^2)$ a non-central F distribution with (M - 1) degrees of freedom in the numerator and M x (N - 1) degrees of freedom in the denominator. If the groups are heterogeneous, a vertical edge exists in the neighborhood of the point (r, s) which causes the group effects $\{a_i\}$ to increase. This increase causes a shift in the mean of ${}_{rs}F$ and the hypothesis H is rejected at a level α. However, since asymptotic normality conditions may require an excessively large sample, the statistic ${}_{rs}f$ can still be used and the decision is made as follows:

$$ {}_{rs}f = \begin{cases} \geq T_o \rightarrow \text{Edge at } (r,s) \\ < T_o \rightarrow \text{No edge at } (r,s) \end{cases} \tag{13} $$

In this case, not only M and N must be properly selected, but also the threshold value T_o in such a way that the total probability of error is minimized. This clearly requires a known sample image within the class under consideration to properly select M, N and T_o through simulation.

The advantage of using quantile and ANOVA techniques over simple ANOVA techniques applied to the original data for the solution of edge detection problems is demonstrated in Figures 1 to 5. Figure 1 represents the uncorrupted reference pattern. To demonstrate the effectiveness of the quantile-ANOVA procedures, the pattern was corrupted with uniform or gaussian noise. Figures 2 to 5 demonstrate the advantages of the new quantile-ANOVA procedures.

Texture edge detection with Gaussian N(0, .5) and N(0, 1) noise simulation is demonstrated in Figures 6 and 7. Here, again, quantile-ANOVA techniques perform very well while the simple ANOVA approach fails to give an outline resembling the original pattern.

One of the disadvantages of one-way ANOVA is that it cannot detect perfectly horizontal edges. This however can be overcome by one-way ANOVA with a reference sample available, or simply by rotating the pattern by 90 degrees and applying one-way ANOVA to it. The processing time for quantile ANOVA techniques was about one third of the direct ANOVA techniques and in a multi-processing environment the ratio should improve in favor of quantile techniques.

The detector parameters for one-way ANOVA, with and without quantile reduction was picked empirically. In particular, M = 2 groups were chosen to limit the neighborhood size and N was allowed to vary from 5 to 15. Best results were obtained for M = 2 and N = 8. The threshold T was also chosen empirically and it was very close to the value corresponding to $\alpha = .01$ if normal distribution is assumed. The simulations demonstrate that quantile-ANOVA techniques perform much better than simple ANOVA techniques for the solution of gray level or texture edge detection. This is especially true if the image is severely corrupted.

Joint Services Technical Advisory Committee
F44620-74-C-0056

L. Kurz and P. Legakis

IBM

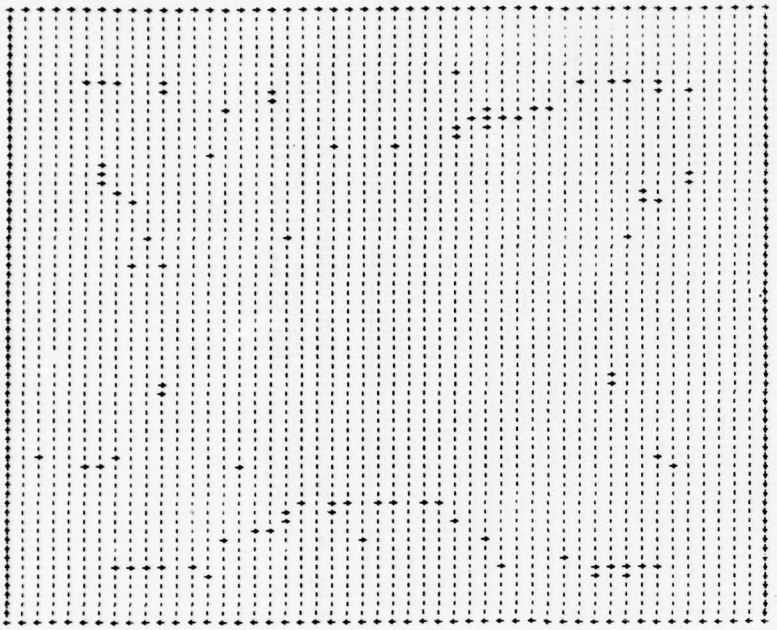


Fig. 2. Edge detection using one-way ANOVA. Reference pattern corrupted with Gaussian N(0, .5) and N(1, .5) noise.

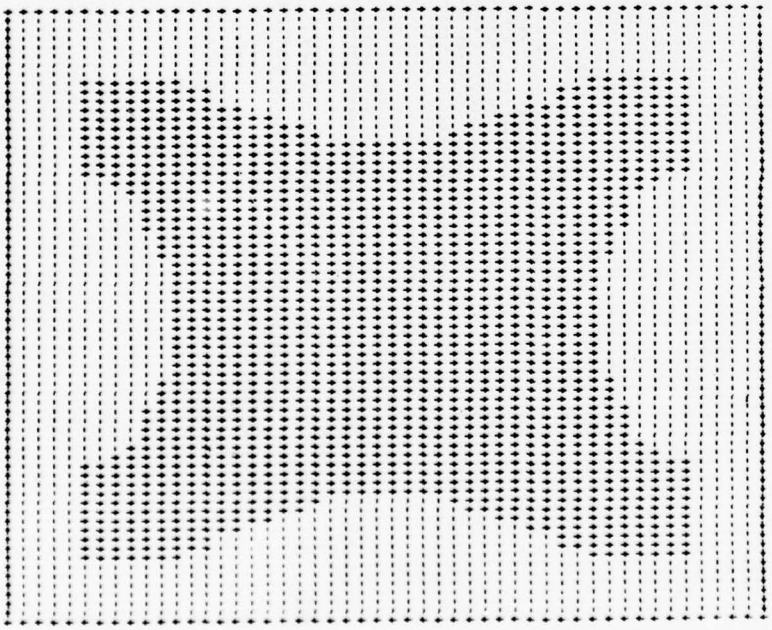


Fig. 1. Uncorrupted binary reference pattern.

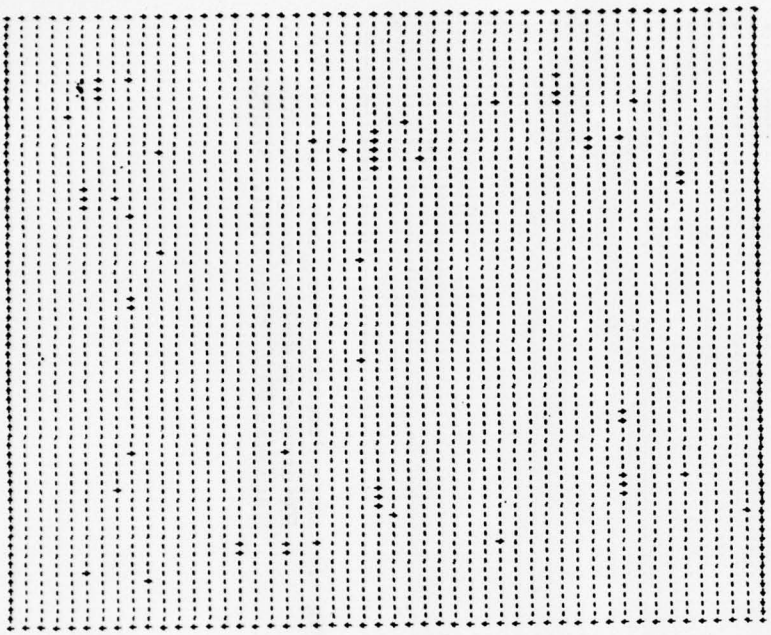


Fig. 4. Edge detection using one-way ANOVA. Reference pattern corrupted with Gaussian N(0,1) and N(1,1) noise.

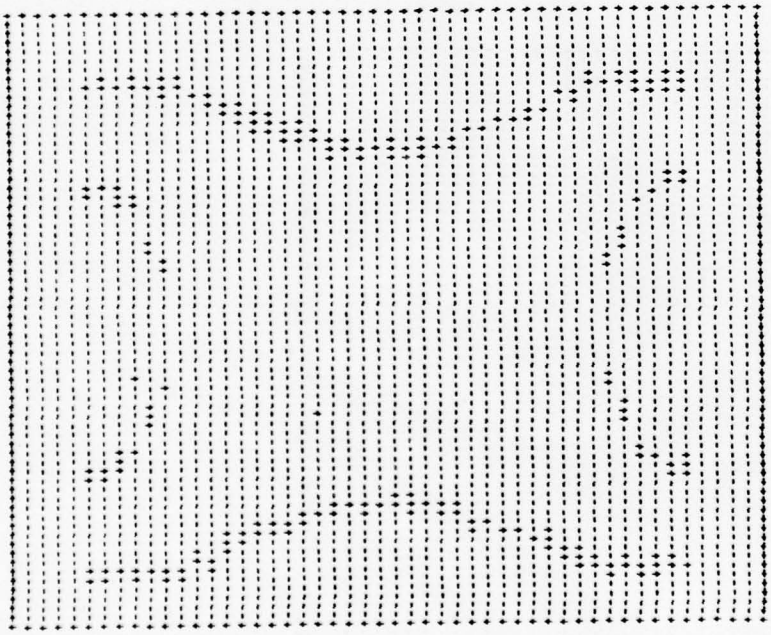


Fig. 3. Edge detection using one-way ANOVA after Quantile estimation. Reference pattern. Corrupted as in Figure 2.

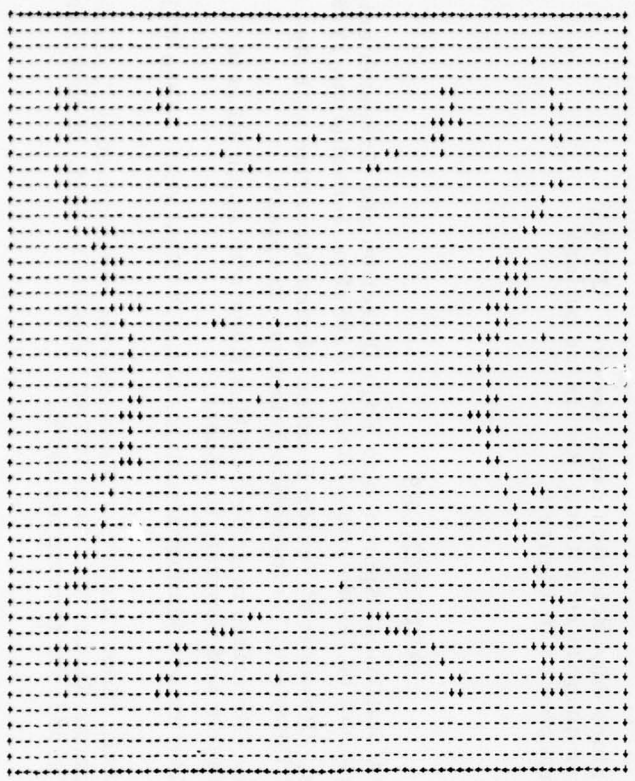


Fig. 5. Edge detection using one-way ANOVA after quantile reduction. Reference pattern corrupted as in Figure 4.

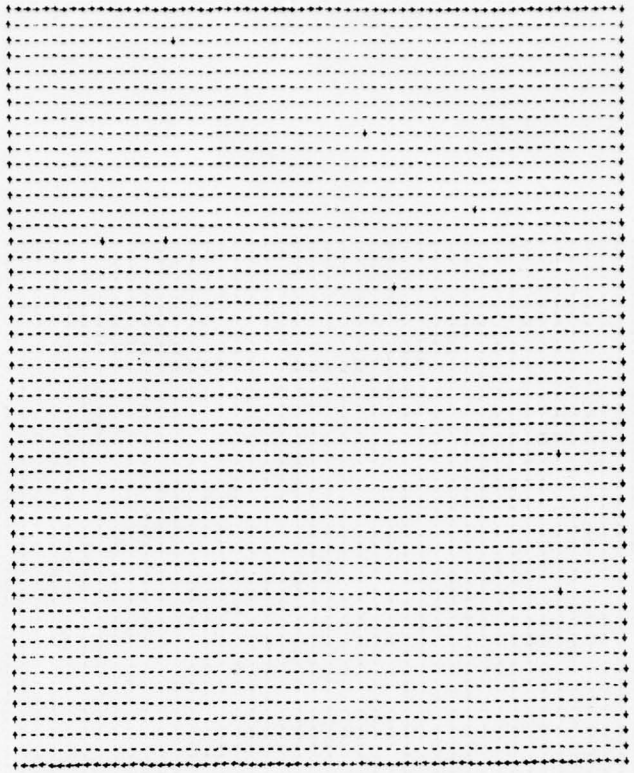


Fig. 6. Texture edge detection using one-way ANOVA. Reference pattern corrupted with Gaussian N(0, .5) and N(0, 1) noise.

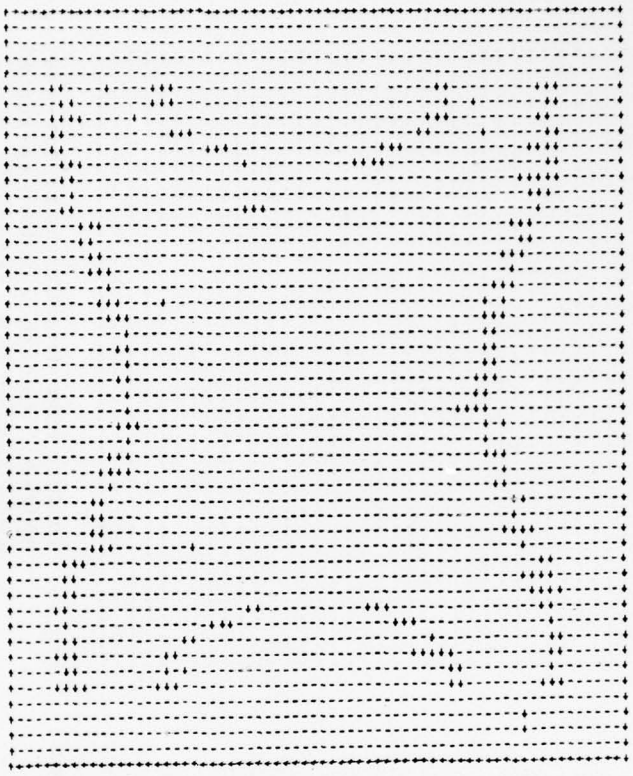


Fig. 7. Texture edge detection using one-way ANOVA after quantile reduction. Reference pattern corrupted as in Figure 6.

REFERENCES

1. J. Aron and L. Kurz, "A Statistical Approach to Edge Detection and Enhancement," 1973 International Symposium on Information Theory, Ashkelon, Israel (June 1973).

2. P. Kersten and L. Kurz, "Robustized Vector Robbins-Monro Algorithm with Applications to M-Interval Detection," Information Sciences, to appear.

3. P. Kersten and L. Kurz, "Bivariate m-Interval Classifiers with Application to Edge Detection," Information and Control, to appear.

4. M. Fisz, "Probability Theory and Mathematical Statistics," (New York: John Wiley, 1963).

5. L. Kurz and C. Mohwinkel, "Some Statistical Enhancement and Extraction Techniques for Tabular Data," Progress Report No. 40 to JSTAC, Polytech. Inst. of New York, Report No. R-452.40-75, 588-593 (November 1975).

GRAY LEVEL DETECTION USING M-INTERVAL OCCUPANCY VECTOR IN CONJUNCTION WITH MANN-WHITNEY STATISTIC

L. Kurz and P. Legakis

In this report, two statistics of the two- and three-sample variety, which are modified versions of the Mann-Whitney statistic and belong to the general class of mixed statistics, are introduced. Unlike in conventional Mann-Whitney statistics of the two- and three-sample type,[1,2] where the processing is performed on the raw data, the new mixed statistic uses as observables the occupancy vector of the m-interval tests.[3] The new statistics preserve the good qualities of both types of tests: power of the standard Mann-Whitney tests and the robustness of the m-interval tests. The mixed tests differ from the original Mann-Whitney statistic in that they require considerably less processing time ($O(m^2)$ vs. $O(n^2)$), $n >> m$ operations) and in that they can be used in shift-in-the-mean and change-of-scale problems, while the standard Mann-Whitney tests on the original observations perform well only in the shift-of-the-mean problems.

A. A Two-Sample Quantile Mann-Whitney Classifier

The availability of a reference sample of n independent and identically distributed (i.i.d.) observations under the hypothesis with distribution, F(x), is assumed. (The influence of dependence on the performance of classifiers discussed in this cahpter, may be studied using the approach suggested by Woinsky and Kurz.[4]) This reference sample is denoted by $Y_n = y_1, y_2, \cdots, y_n$. The received sample of n i.i.d. observations to be classified is denoted by $X_n = x_1, x_2, \cdots, x_n$. From the reference sample a vector $\vec{a} = (a_0, a_1, \cdots, a_m)$ corresponding to the m-interval quantile partitioning of F(x) is formed such that

$$\text{Prob}[a_{i-1} < y_j \leq a_i] = F_Y(a_i) - F_Y(a_{i-1}) = p_i$$

$$i = 1, 2, \cdots, m; \quad j = 1, 2, \cdots, n \tag{1}$$

$$a_o > -\infty; \quad a_m < \infty; \quad p_i = \frac{1}{m}$$

Let n_i be a set of numbers defined as

$$n_i = \text{Card}\,\{y_j : a_{i-1} < y_j \leq a_i\}$$

$$i = 1, 2, \cdots, m; \quad j = 1, 2, \cdots, n \tag{2}$$

The classification problem is to decide whether X_n and Y_n come from the same distribution or not. With a_i as defined by Eq. (1), a new sequence ℓ_i is generated such that

$$\ell_i = \text{Card}\ \{x_j : a_{i-1} \leq x_j < a_i \tag{3}$$

$$i = 1,2,\cdots,m; \quad j = 1,2,\cdots,n$$

The Mann-Whitney statistic, U, based on n_i and ℓ_j, the occupancy vectors, is now generated by

$$U = \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} u(n_i - \ell_j) \tag{4}$$

where

$$u(\cdot) = \begin{cases} 1 & \text{if} \quad n_i \geq \ell_j \\ 0 & \text{if} \quad n_i < \ell_j \end{cases} \tag{5}$$

The U statistic counts the number of times n_i is greater than ℓ_j and on the basis of that count the classifier makes a decision as to whether X_n comes from the distribution, F(x), under the hypothesis, or the distribution, G(x), under the alternative. This formulation is distribution-free not only because quantile information which may be obtained from the data rather than exact knowledge of the appropriate distribution functions is involved but also because the U statistic is based on the relative ordering of the n_i and ℓ_j and not on any known distributions.

B. Distribution of the Two-Sample Quantile Mann-Whitney Statistic Under the Hypothesis (H) and the Alternative (K)

Mann and Whitney[5] have shown that U is asymptotically normally distributed under both the hypothesis and the alternative, provided that the n_i and ℓ_j are i.i.d.r.v. It is known[6] that the n_i are multinomially distributed

$$u(n_1,n_2,\cdots,n_m) = \frac{n!}{\prod_{i=1}^{m} n_i} \prod_{i=1}^{m} p_i^{n_i}$$

subject to the constraints

$$\sum_{i=1}^{m} n_i = n \quad \text{and} \quad \sum_{i=1}^{m} p_i = n$$

Also

$$E[n_i] = np_i; \quad \text{Var}[n_i] = np_i(1 - p_i); \quad \text{Cov}(n_i, n_j) = -n_i p_i p_j$$

Extensive simulation studies under the same signal conditions and using Kendall's τ coefficient for independence, show that this dependence is very weak. In fact, simulation

results show that the n_i are independent at least at the $\alpha = .01$ level and, therefore, the moments of U in the form derived by Mann-Whitney can be used. The first and second moments of the U-statistic are given by

Under H:

$$E_H[U] = m_{U/H} = .5$$

$$\mathrm{Var}_H[U] = \sigma^2_{U/H} = \frac{1}{12m^2}(2m + 1)$$

Under K:

$$E_K[U] = m_{U/K} = .5 - \lambda$$

$$\mathrm{Var}_K[U] = \sigma^2_{U/K} = \frac{1}{12m^2} - \frac{\lambda^2}{m^2}(2m+1) + \frac{(m-1)}{m^2}(2\lambda - \varepsilon_1 - \varepsilon_2)$$

where

$$\lambda = \frac{1}{2} - \int_{-\infty}^{\infty} G(x)\, dF(x)$$

$$\varepsilon_1 = \frac{1}{3} - \int_{-\infty}^{\infty} G^2(x)\, dF(x)$$

$$\varepsilon_2 = \frac{1}{3} - \int_{-\infty}^{\infty} (1 - F(x))^2\, dG(x)$$

Since the first and second moments of a normal distribution completely define the distribution and since the Mann-Whitney statistic is asymptotically normal, the asymptotic distributions under the hypothesis and the alternative are known.

Under uniform noise conditions and shift-in-the-mean alternatives the first and second order moments of U can be easily specified by calculating λ, ε_1 and ε_2. Namely

Under H:

$$F(x) = G(x) = xu(x) - (x - 1)\, u(x - 1)$$

Under K:

$$F(x) = xu(x) - (x - 1)\, u(x - 1)$$

and

$$G(x) = F(x - c) = (x - c)\, u(x - c) - (x - c - 1)\, u(x - c - 1)$$

where u(x) is the unit step function and c is a constant shift parameter. Thus

$$\lambda = \frac{1}{2} - \int_{-\infty}^{\infty} G(x)\, dF(x) = \frac{1}{2} - \int_{c}^{1} (x - c)\, dx = \frac{1}{2}(1 - (1 - c)^2)$$

$$\varepsilon_1 = \frac{1}{3} - \int_{-\infty}^{\infty} G^2(x)\, dF(x) = \frac{1}{3} - (1 - (1 - c)^3)$$

$$\varepsilon_2 = \frac{1}{3} - \int_{-\infty}^{\infty} dG(x) + 2\int_{-\infty}^{\infty} F(x)\, dG(x) - \int_{-\infty}^{\infty} F(x)^2\, dG(x) = (\frac{c^2}{3} - c + 1)c$$

Direct substitution of the expression for λ, ε_1 and ε_2 gives the expression for the moments of U in terms of the partition size and the shift parameter c.

C. The Two-Sample Quantile Classification Procedure

The two-sample quantile classification procedure can be formulated as follows:

(1) From the reference sample Y_n compute the quantile vector $\vec{a} = (a_0, a_1, \cdots, a_m)$ which partitions the probability space under the hypothesis into m equiprobable segments

(2) From the same reference sample compute the n_i which denote how many observations from the reference sample fall into each of the cells defined by $\vec{a}$

(3) From the sample to be classified compute the ℓ_i which denote how many observations from X_n fell into each of the cells denoted by $\vec{a}$

(4) On the basis of n_i and ℓ_j compute the U-statistic as indicated by Equation (4)

(5) Compare the value of U with the threshold U_o which minimizes the average probability of error. (Other criteria for selection of U such as Bayes, minimax, etc., are also applicable.)

(6) Decide that the hypothesis is true if $U \geq U_o$ or that the alternative is true if $U < U_o$.

The average probability of incorrect classification or average probability of error, P_e, is given by

$$P_e = \alpha p_H + \beta p_K \tag{6}$$

where

p_H a priori probability of H
p_K a priori probability of K
α the type-I error
β the type-II error

Equation (6) may be written explicitly as

$$P_e = p_H \int_{-\infty}^{U_o} f_H(u)\,du + p_K \int_{U_o}^{\infty} f_K(u)\,du \tag{7}$$

Since the a priori probabilities p_H and p_K are not known, it is customary to assume that they are equally likely

$$p_H = p_K = \frac{1}{2}$$

It can be shown that the optimum threshold is

$$U_o = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}$$

where

$$A = \sigma^2_{U/H} - \sigma^2_{U/K}$$

$$B = 2(m_{U/H}\sigma^2_{U/K} - m_{U/K}\sigma^2_{U/H})$$

$$C = (m_{U/K})^2 - (m_{U/H}\sigma_{U/K})^2 - 2\sigma^2_{U/H}\sigma^2_{U/K}\ln(\sigma_{U/H}/\sigma_{U/K})$$

The proper choice between the two values of U_o is obvious.

D. Three-Sample Quantile Mann-Whitney Statistic

It is intuitively appealing that if in addition to the reference sample which is taken from the distribution under the hypothesis, a sample taken from the distribution under the alternative was available at the classifier, a new test statistic could be formulated which would take advantage of the additional information resulting in improved performance of the classifier. The three-sample quantile classification problem can be stated as follows.

Let $\vec{Y}_n = (y_1, y_2, \cdots, y_n)$ and $\vec{Z}_n = (z_1, z_2, \cdots, z_n)$ be two vectors consisting of n i.i.d.r.v., each, and taken from a hypothesis distribution H and an alternative distribution K, respectively, i.e., $Y_n \varepsilon F_Y(\cdot)$ and $Z_n \varepsilon F_Z(\cdot)$. Let $\vec{a} = (a_o, \cdots, a_m)$ and $\vec{\alpha} = (\alpha_o, \cdots, \alpha_m)$ be the corresponding m-interval equiprobable quantile partitionings under the hypothesis and the alternative distributions respectively. Namely,

$$\text{Prob}[a_{i-1} < y_j \leq a_i] = p_i$$

$$\text{Prob}[\alpha_{i-1} < z_j \leq \alpha_i] = p_i$$

$$i = 1, 2, \cdots, m; \quad j = 1, 2, \cdots, n$$

$$a_o, a_o > -\infty; \quad a_m, a_m < +\infty; \quad \sum_i^m p_i = 1$$

Let $X_n = (x_1, x_2, \cdots, x_n)$ be a vector of observations that need be classified as coming either from H or K, and let $\vec{\ell} = (\ell_1, \ell_2, \cdots, \ell_m)$ and $\vec{w} = (w_1, \cdots, w_m)$ be two vectors such that

$$\ell_i = \text{Card } \{x_j : a_{i-1} < x_j \leq a_i\}$$

$$w_i = \text{Card } \{x_j : a_{i-1} < x_j \leq a_i\}$$

$$i = 1, 2, \cdots, m; \quad j = 1, 2, \cdots, n$$

The Mann-Whitney quantile statistic can now be formulated as follows:

$$U = \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} u(\ell_i - w_j)$$

where

$$u(\ell_i - w_j) = \begin{cases} 1 & \text{if} \quad \ell_i \geq w_j \\ 0 & \text{if} \quad \ell_i < w_j \end{cases}$$

Note that the formulations of the two- and three-sample tests are similar but with an important difference. While in the two-sample case the Mann-Whitney statistic measures some kind of a relative distance of the observable from the hypothesis, in the three-sample case it measures the relative difference between the hypothesis and the alternative causing the improvement in performance of the three-versus the two-sample case. Also, if the formulation of the two-sample Mann-Whitney statistic based on observations X_n and Y_n requires $O(n^2)$ operations, the formulation of the three-sample statistic requires $2O(n^2)$ operations. Such, however, is not the case in the formulation of the two- and three-sample quantile Mann-Whitney statistics because they require the same number, $O(m^2)$ operations once the quantile occupancy vectors $\vec{\ell} = (\ell_1, \ell_2, \cdots, \ell_m)$ and $\vec{w} = (w_1, w_2, \cdots, w_m)$ are available, and in usual applications $m << n$. It can be shown that the optimum threshold for the three-sample test is

$$U_o = \frac{1}{2} + \sigma^2_{U/H} \ln \frac{P_K}{P_H}$$

and the corresponding probability of error is

$$P_e \cong \frac{1}{2} \left\{ 1 - \frac{\delta}{\sqrt{2\pi}\, \sigma_{U/H}} \right\}$$

where δ is a small signal approximation to the parameter

$$\lambda = \int_{-\infty}^{\infty} (F_y(\xi) - F_z(\xi))\, dF_y(\xi)$$

E. Simulation Results

In order to subjectively evaluate the two- and three-sample quantile Mann-Whitney statistics, they were applied on a pattern which is based on the binary pattern of Figure 1. The symbols '$\ddagger$' and 'X' are used to indicate binary zeros and ones, respectively. A pictorial view of what noise of a certain type does to a binary pattern is given in Figures 2 through 4. For example, Fig. 3 results from Fig. 1 when '$\ddagger$' and 'X' are replaced by observations from N(0, 1) and N(1, 1), respectively, and a threshold of .5 is applied. To indicate how well the two-sample statistic performs under the same conditions as those of Fig. 3, reference is made to Figure 5. The three-sample statistic performs considerably better (not shown because it looks almost like the reference pattern). Figure 6 illustrates the performance of the three-sample statistic under severe [N(0, 1) and N(.5, 1)] noise conditions. Although the detected pattern is not as good as Fig. 1, it is considerably better than Figure 4. Figure 7 illustrates the performance of the three-sample statistic under change-of-scale conditions. It is to be noted that the original Mann-Whitney statistic takes considerably longer to process for approximately the same performance under severe noise conditions and that it cannot be applied to change-of-scale problems.

Joint Services Technical Advisory Committee
F44620-74-C-0056

L. Kurz and P. Legakis

IBM

REFERENCES

1. M.N. Woinsky and L. Kurz, "An Efficient Nonparametric Detector Based on a Three-Sample Classification Model," IEEE Trans. on Information Tneory, IT-14, 744-752 (September 1968).
2. M.N. Woinsky and L. Kurz, "Sequential Nonparametric Two-Way Classification with Prescribed Maximum Asymptotic Error Probabilities," Annals of Math. Stat. 40, No. 2, 445-455 (April 1969).
3. Y.C. Ching and L. Kurz, "Nonparametric Detectors Based on m-Interval Partitioning," IEEE Trans. on Information Theory, IT-18, No. w, 25-1258 (March 1972).
4. M.N. Woinsky and L. Kurz, "Nonparametric Detection Using Dependent Samples," IEEE Trans. on Information Theory, IT-16, No. 3, 355-358 (May 1970).
5. H.B. Mann and D.R. Whitney, "On a Test of Whether One or Two Random Variables Is Stochastically Larger Than the Other," Ann. of Math. Stat., 18, 50-60 (1947).
6. A.M. Kshirsagav, "Multivariate Analysis," (New York: Marcel Decker Inc., 1972).

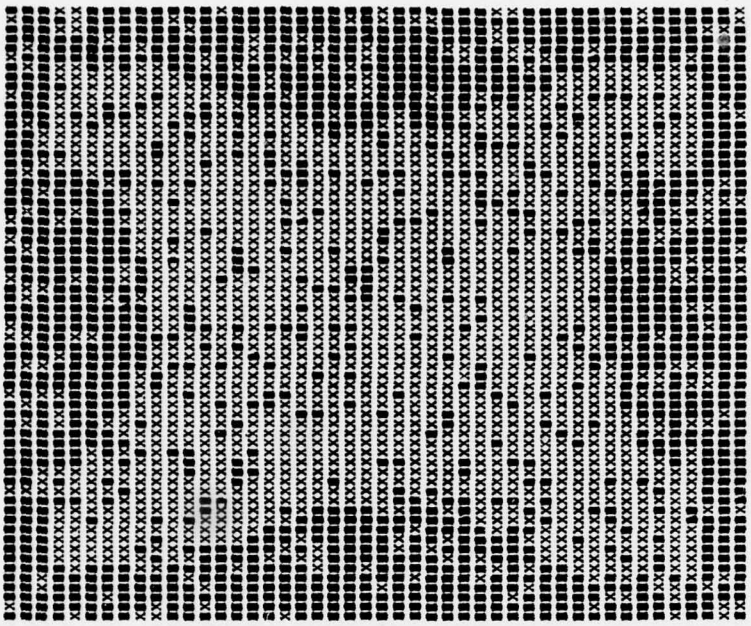


Fig. 2. Reference pattern corrupted with gaussian N(0, .5) and N(1, .5) noise.

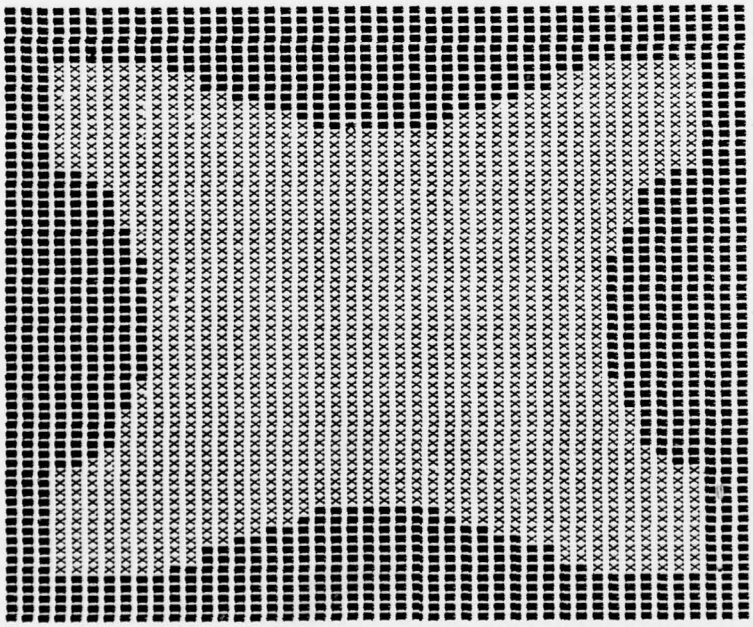


Fig. 1. 2-gray level reference pattern.

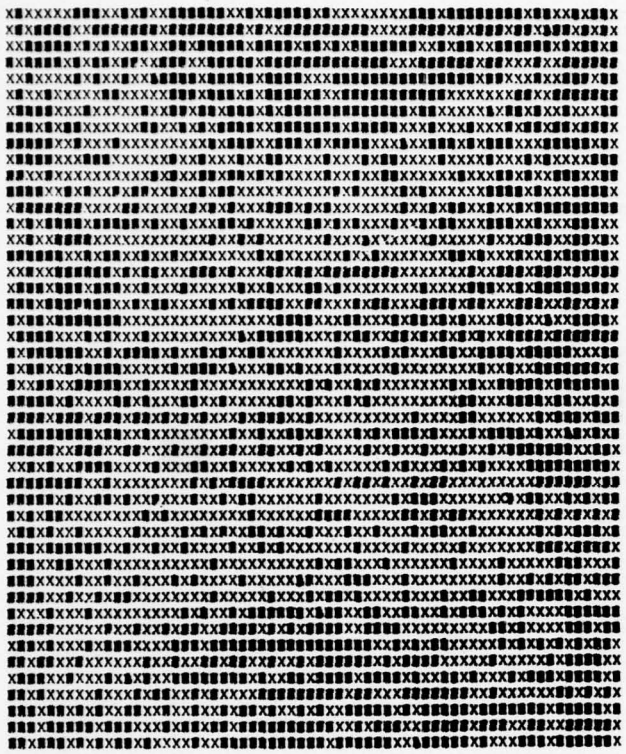


Fig. 3. Reference pattern corrupted with gaussian N(0, 1) and N(1, 1) noise.

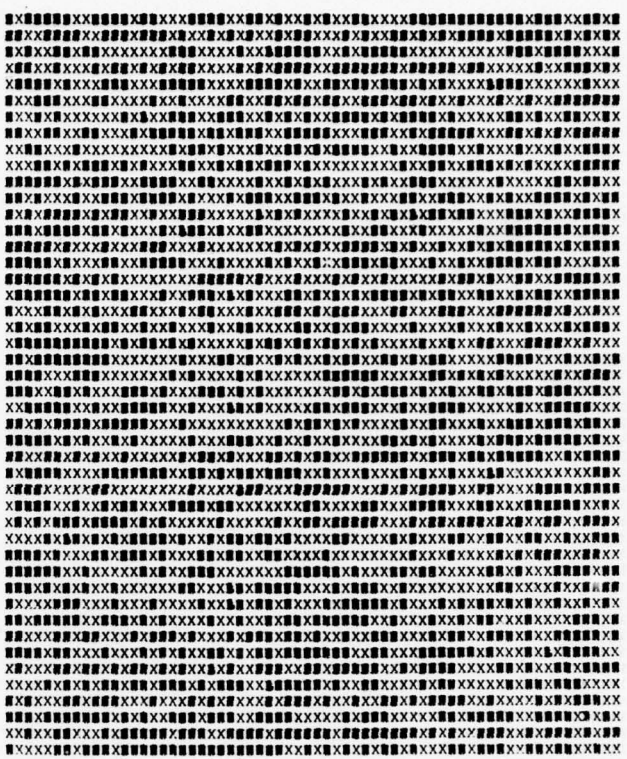


Fig. 4. Reference pattern corrupted with gaussian N(0, 1) and N(.5, 1) noise.

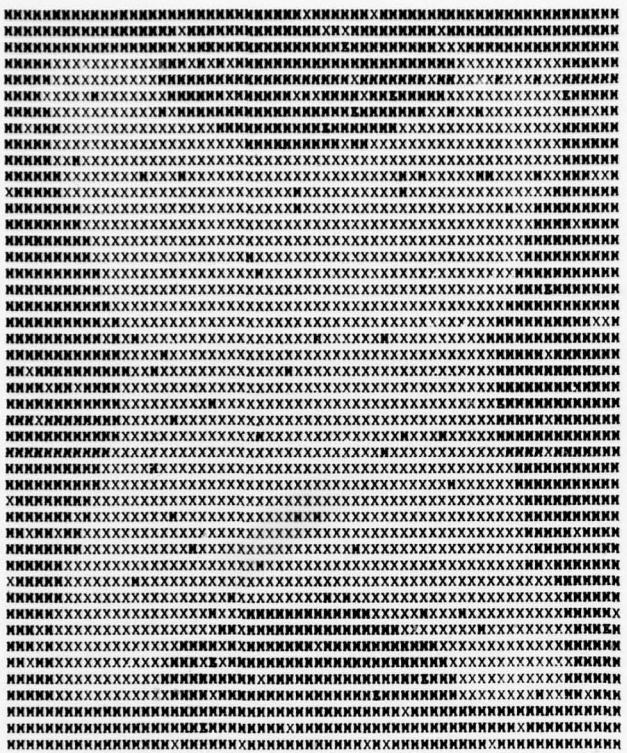


Fig. 5. Gray level detection using the two-sample Mann-Whitney statistic based on quantile occupancy vectors. Reference pattern corrupted with gaussian N(0, 1) and N(0, .5) noise.

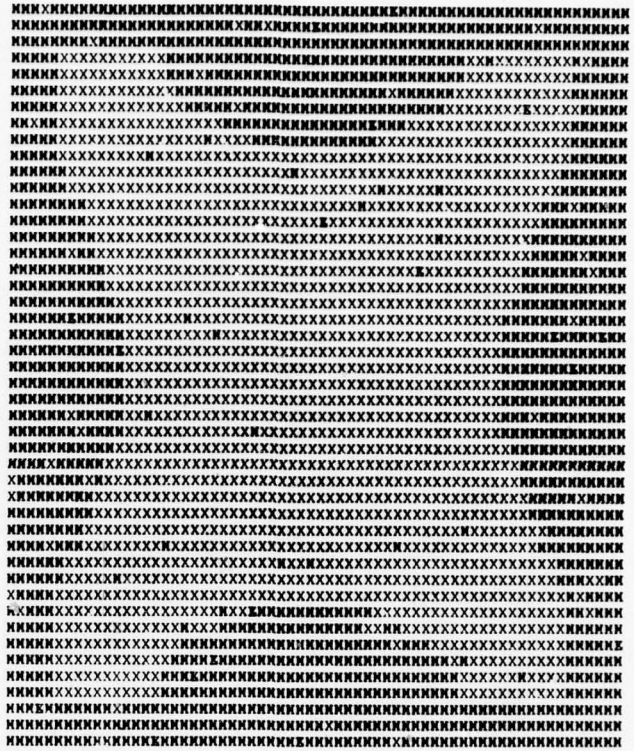


Fig. 6. Texture detection using the three-sample Mann-Whitney statistic based on quantile occupancy vectors. Reference pattern corrupted with gaussian N(0, .5) and N(0, 1) noise.

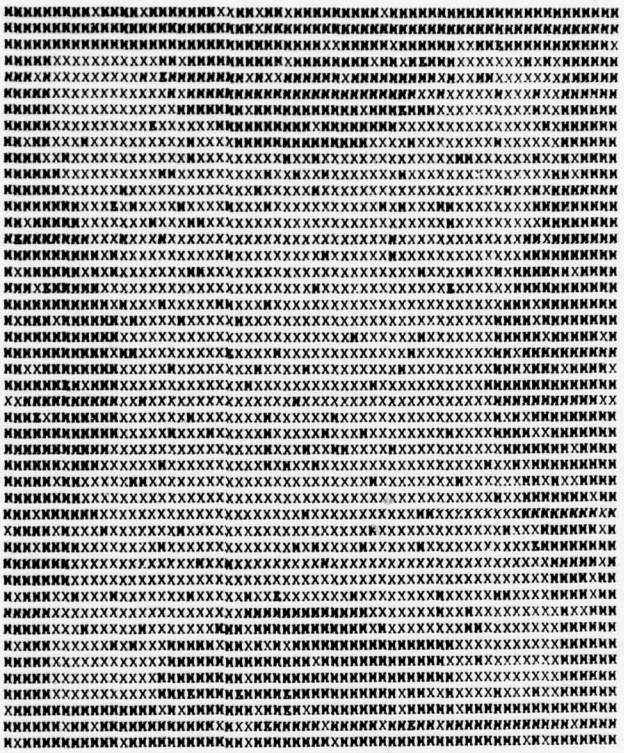


Fig. 7. Gray level detection using the three-sample Mann-Whitney statistic based on quantile occupancy vectors. Reference pattern corrupted with gaussian N(0, 1) and N(.5, 1) noise.

DETERIORATION IN PERFORMANCE OF A DUOBINARY PAM SYSTEM BY INTRODUCTION OF NONLINEARITIES

L. Kurz and M. Wernicki

In this report, the effect of memoryless nonlinearities on intersymbol interference in a duobinary PAM system is considered. As was shown in Refs. 1 and 2, nonlinearities are introduced in a PAM system to improve overall performance if the sources of degradation include not only intersymbol interference and gaussian noise but also impulsive noise. It was demonstrated that the influence of nonlinearities natural to each system or introduced to suppress impulsive noise creates problems in controlling intersymbol interference.

The analysis in Refs. 1 and 2 was concentrated on the third order polynomial nonlinearity as a best representative of the general class of nonlinearities which approximate natural or near-optimum for impulsive noise nonlinearities. This model is used here, too. The analysis in these references which concentrated on standard PAM systems is extended to PAM systems with duobinary signalling.[3,4] Using the prolate spheroidal wave function (PSWF) in a series expansion, bounds are found on intersymbol interference introduced by the nonlinearity. Related problems of elimination of intersymbol interference caused by the nonlinearity are also considered. The importance of the duobinary system, which has the rate capability of a quaternary system but requires only slightly more complicated circuitry than a standard binary PAM system, makes the study of it particularly useful.

A. Bound on the Intersymbol Interference Caused by the Third Power Nonlinearity -- PSWF Series Approach

The duobinary signal is of the form

$$s(t) = \frac{4}{\pi} \frac{\cos \pi \frac{t}{T}}{1 - 4\frac{t^2}{T^2}} \tag{1}$$

with Fourier transform

$$S(\omega) = 2T \cos\left(\frac{T}{2}\omega\right) \qquad |\omega| \leq \frac{\pi}{T} \tag{2}$$

$$= 0 \qquad \text{otherwise}$$

where the notation of Ref. 4 has been used. The transfer functions of the corresponding transmitting and receiving filters are

$$G_T(\omega) = G_R(\omega) = \left(2T \cos \frac{T}{2}\omega\right)^{1/2}, \qquad |\omega| \leq \frac{\pi}{T} \tag{3}$$

$$= 0 \qquad \text{otherwise}$$

At the output of the nonlinearity $f(x) = a_o + a_3 x^3$ with proper truncation (see Ref. 1), the only term which will contribute to intersymbol interference will be the cubic term

$$y_i(t) = \left(\frac{4}{\pi} \frac{\cos \frac{\pi t}{T}}{1 - 4\frac{t^2}{T^2}}\right)^3 \tag{4}$$

Consider the expansion

$$\hat{y}_i(t) = \sum_{n=o}^{N} a_{2n}\, \theta_{2n}(c,t) \qquad |t| \le \frac{\pi}{T}$$

$$= 0 \qquad \text{otherwise} \tag{5}$$

where $\{\theta_{2n}(c,t)\}$ are the prolate spheroidal wave functions generated by the integral equation

$$\lambda_n(c)\, \theta_n(c,v) = \int_{-\pi/T}^{\pi/T} \frac{\sin T(v-x)}{(v-x)}\, \theta_n(c,x)\, dx \qquad n = 0, 1, 2, \cdots \tag{6}$$

These functions satisfy

$$\int_{-\infty}^{\infty} \theta_n(c,v)\, \theta_m(c,v)\, dv = \delta_{mn} \tag{7}$$

and

$$\int_{-\pi/T}^{\pi/T} \theta_n(c,v)\, \theta_m(c,v)\, dv = \delta_{mn}\, \lambda_n(c) \tag{8}$$

where δ_{mn} is the Kronecker delta. In addition,

$$2(j)^n \sqrt{\lambda_n(c)}\, \theta_n(c,\omega) = \int_{-\pi/T}^{\pi/T} e^{j\omega t}\, \theta_n(c,t)\, dt \tag{9}$$

In view of the properties of PSWF, it can be shown that the coefficients in Eq. (5) are

$$a_{2n} = \frac{(-1)^n}{2[\lambda_{2n}(c)]^{3/2}} \int_{-\pi/T}^{\pi/T} Y_i(\omega)\, \theta_{2n}(c,\omega)\, d\omega \tag{10}$$

where

$$Y_i(\omega) = \mathcal{F}[y_i(t)] \cong \hat{Y}_i(\omega) = \sum_{k=o}^{K} \frac{\omega^{2k}}{(2k)!} (-1)^k s_{2k} \tag{11}$$

and s_{2k} are the moments of $y_i(t)$. Using the mean-square error criterion,

$$\varepsilon^2 = \frac{\int_{-\pi/T}^{\pi/T} [y_i(t) - \hat{y}_i(t)]^2 \, dt}{\int_{-\pi/T}^{\pi/T} y_i^2(t) \, dt} \tag{12}$$

one needs to find a bound on the numerator of Equation (12). This numerator may be written in the form

$$\int_{-\pi/T}^{\pi/T} [y_i(t) - \hat{y}_i(t)]^2 \, dt = \sum_{n=o}^{N} (a_{2n} - \hat{a}_{2n})^2 + \sum_{n=N+1}^{\infty} a_{2n}^2 \tag{13}$$

where the orthogonality property of the PSWF on $[-\frac{\pi}{T}, \frac{\pi}{T}]$ has been used. But

$$(a_{2n} - \hat{a}_{2n}) = \frac{1}{\lambda_{2n}^3(c)} \left\{ \int_{-\pi/T}^{\pi/T} [Y_i(\omega) - \hat{Y}_i(\omega)] \, \theta_{2n}(c, \omega) \, dv \right\}^2 \tag{14}$$

and, applying Schwarz' inequality to Eq. (14), yields

$$(a_{2n} - \hat{a}_{2n}) \leq \frac{1}{\lambda_{2n}^3(c)} \int_{-\pi/T}^{\pi/T} [Y_i(\omega) - \hat{Y}_i(\omega)]^2 \, d\omega \int_{-\pi/T}^{\pi/T} \theta_{2n}^2(c, \omega) \, d\omega$$

$$= \frac{1}{\lambda_{2n}^3(c)} \int_{-\pi/T}^{\pi/T} [Y_i(\omega) - \hat{Y}_i(\omega)]^2 \, dv \tag{15}$$

Since

$$[Y_i(\omega) - \hat{Y}_i(\omega)] \leq \frac{\omega^{(2k+2)}}{(2k+2)!} s_{2k+2} \tag{16}$$

where use has been made of Eq. (11)

$$(a_{2n} - \hat{a}_{2n}) \leq \frac{c^{4k+6}}{[(2k+3)!]^2 (4k+5)} \frac{1}{\lambda_{2n}^2(c)} \tag{17}$$

and, introducing a mean-square error $\varepsilon_{\omega_o}^2$ for $Y_i(\omega)$ of the same form as Eq. (12), results in

$$\int_{-\pi/T}^{\pi/T} [y_i(t) - \sum_{n=o}^{c/\pi} a_n \theta_n(c, t)]^2 \, dt \leq 12 \, \varepsilon_{\omega_o}^2 \int_{-\pi/T}^{\pi/T} y_i^2(t) \, dt \tag{18}$$

which leads to

AD-A034 764 POLYTECHNIC INST OF NEW YORK BROOKLYN MICROWAVE RESE--ETC F/G 20/3
PROGRESS REPORT NUMBER 41, 15 SEPTEMBER 1975 THROUGH 14 SEPTEMB--ETC(U)
NOV 76 A A OLINER F44620-74-C-0056

UNCLASSIFIED POLY-MRI-452.41-76 NL

5 OF 7
AD
A034764



$$\sum_{n=\frac{c}{\pi}+1}^{\infty} a_{2n}^2 \leq 12\,\varepsilon_{\omega_o}^2 \int_{-\pi/T}^{\pi/T} y_i(t)\,dt \quad , \quad \frac{c}{\pi} = N \tag{19}$$

The estimation error, ε^2, is then bounded by

$$\varepsilon^2 \leq \frac{c^{4k+6}}{[(2k+3)!]^2\,(4k+5)} \frac{\sum_{n=o}^{c/\pi} \frac{1}{\lambda_{2n}^2(c)}}{\int_{-\pi/T}^{\pi/T} y_i^2(t)\,dt} + 12\,\varepsilon_{\omega_o}^2 \tag{20}$$

A pessimistic estimate of this error is obtained by using an upper bound for $\varepsilon_{\omega_o}^2$ and a lower bound for $\int_{-\pi/T}^{\pi/T} y_i^2(t)\,dt$ in Equation (20). Typically $\varepsilon_{\omega_o}^2 < .01$ and using lower bound on $\int_{-\pi/T}^{\pi/T} y_i^2(t)\,dt$ results for $c = 1$ and $k = 1$ in $\varepsilon^2 \leq 3.46 \times 10^{-2}$.

B. Some Comments on Elimination and Boundedness of Intersymbol Interference

Following the same reasoning as in Ref. 2, it can be shown that it is not always possible to eliminate intersymbol interference by linear filtering when a memoryless nonlinearity is present. As a matter of fact, the theorems given in the reference are still valid. It remains to demonstrate that the boundedness of the intersymbol interference is guaranteed.

Consider the cubic term of the output of the nonlinearity followed by a linear filter with impulsive response, h(t). The output of the filter is obtained by convolving $y_i(t)$ and h(t). At any sampling instant $t = nT$, the output of the linear filter is

$$z(0) = \int_{-NT}^{NT} y_i(t)\,h(-t)\,dt \tag{21}$$

Applying Schwarz' inequality

$$[z(0)]^2 \leq \int_{-NT}^{NT} y_i^2(t)\,dt \int_{-\infty}^{\infty} h^2(-t)\,dt \tag{22}$$

There exists an M such that

$$|x^M| \geq |y_i(t)| \tag{23}$$

or

$$\int_{-NT}^{NT} y_i^2(t)\,dt \leq \int_{-NT}^{NT} \left[\frac{4}{\pi} \frac{\cos\frac{\pi t}{T}}{1 - \frac{4t^2}{T^2}}\right]^{2M} dt \tag{24}$$

from which it follows that

$$\int_{-NT}^{NT} \left| \frac{4}{\pi} \frac{\cos \frac{\pi t}{T}}{1 - \frac{4t^2}{T^2}} \right|^{2M} dt \leq \int_{-NT}^{NT} \frac{4}{\pi} \left| \frac{1}{1 - \frac{4t^2}{T^2}} \right|^{2M} dt \tag{25}$$

From Eqs. (22) and (25), we conclude that intersymbol interference is bounded provided that h(t) is square-integrable. If the nonlinearity is more general than cubic, the boundedness of the intersymbol interference is guaranteed if the new $y_i(t)$ satisfies Equation (23).

Joint Services Technical Advisory Committee
F44620-74-C-0056

L. Kurz and M. Wernicki

REFERENCES

1. G. Soloway and L. Kurz, "Performance of Suboptimal Nonlinear Detectors in Gaussian and Impulsive Noise," Progress Report No. 40 to JSTAC, Polytech. Inst. of New York, Report No. R-452.40-74, 386-395 (November 1975).
2. L. Kurz and G. Soloway, "The Effect of Nonlinearities on Suppression of Intersymbol Interference," Progress Report No. 40 to JSTAC, Polytech. Inst. of New York, Report No. R-452.40-74, 402-406 (November 1975).
3. A. Lender, "The Duobinary Technique for High-Speed Data Transmission," Trans. on Communication Electronics, 214-218 (May 1963).
4. R.W. Lucky, J. Satz and E.J. Weldon, "Principles of Data Communications," (New York: McGraw Hill, 1968).

BIVARIATE DEPENDENT SEQUENTIAL PARTITION DETECTORS

R.F. Dwyer and L. Kurz

The purpose of this report is to extend the theory of partition detectors[1] to include bivariate dependent sequential partition detection. The extension is of particular use in picture, radar and underwater sounding data processing. General formulation of the problem is followed by application to the shift-of-the-mean and Lehmann alternatives model.

A. Structure of Bivariate Dependent Sequential Partition Detectors (BSPD)

Let $z_i = (x_i, y_i)$, $i = 1, 2, \cdots, n$ be an i.i.d. sample from a stationary bivariate random process with c.d.f. $F(x,y)$. Define $\vec{a} = (a_o = -\infty, a_1, \cdots, a_{m-1}, a_m = +\infty)$ as an ordered vector of quantiles, estimated with independent samples for c.d.f $F(x)$. Assume $F(x,y) = F(x)\,F(y)$ when independent, and a partitions the space into m^2 mutually exclusive cells in R_2.

Define n_{kj} as the number of observations falling into the two-dimensional cell (k,j),

$$I_{kj}\left[a_{k-1} < x_i \leq a_k,\ a_{j-1} < y_i \leq a_j\right];\ i = 1,2,\cdots,n;\ k = 1,2,\cdots,m,\ j = 1,2,\cdots,m]$$

and

$$P_{Lkj} = P_{Lkj}[x_i, y_i] = P_L[a_{k-1} < x_i \leq a_k,\ b_{j-1} < y_i \leq b_j],\ L = 0,1$$

as the probability of the joint event where

$$\sum_k^m \sum_j^m P_{Lkj} = 1 \quad \text{and} \quad \sum_k^m \sum_j^m n_{kj} = n$$

From a direct generalization of the one-dimensional m-interval case, the variable n_{kj} are multinomially distributed (see Ref. 2, pp. 168), with

$$P_L(\vec{n}) = P(n_{11}, n_{12}, \cdots, n_{1m}, \cdots, n_{mm})$$

$$= \frac{n!}{n_{11}!n_{12}! \cdots n_{kj}! \cdots n_{mm}!} \prod_k^m \prod_j^m P_{Lkj}^{\,n_{kj}}$$

Define

$$T_n = \ell n\left[\frac{P_1(\vec{n})}{P_o(\vec{n})}\right] = \sum_k^m \sum_j^m n_{kj}\, \ell n\left[\frac{P_1(a_{k-1} < x_i \leq a_k,\ a_{j-1} < y_i \leq a_j)}{P_o(a_{k-1} < x_i \leq a_k,\ a_{j-1} < y_i \leq a_j)}\right] \tag{1}$$

as the loglikelihood ratio for $(z_1, z_2, \cdots, z_n)$ under H_o and H_1.

Let

$$n_{kj} = \sum_{i}^{n} n_{ikj}$$

where

$$n_{ikj} = \begin{cases} 0 & \text{if } z_i \notin I_{kj} = (a_{k-1}, a_k; a_{j-1}, a_j) \\ 1 & \text{if } z_i \in I_{kj} \end{cases}$$

Then,

$$T_n = \sum_{i}^{n} \sum_{k}^{m} \sum_{j}^{m} n_{ikj} b_{kj} \tag{2}$$

where

$$b_{kj} = \ell n \left[\frac{P_1(z_i)}{P_o(z_i)} \right]$$

Equation (2) represents the bivariate sequential partition detector (BSPD). For the data sequence $z_1, z_2, \cdots, z_n$, the test T_n, continues sampling if $b < T_n < a$ and terminates with acceptance of H_o or H_1 if $T_n \leq b$ or $T_n \geq a$, respectively.

Define the random variable,

$$T_i = \sum_{k}^{m} \sum_{j}^{m} b_{kj} n_{ikj} \quad , \tag{3}$$

with moment generating function

$$\phi(t) = E[e^{t T_i}]$$

Following the usual procedure for sequential tests as in Eq. (3), a value of $t = t_o \neq 0$ can be found so that $\phi(t) = 1$, and

$$\lim_{n \to \infty} t_o = -2 \frac{E(T_i)}{\sigma^2_{T_L}}$$

The average sample number for BSPD is

$$ASNB = \frac{1}{E(T_i)} \frac{b(e^{t_o a} - 1) + a(1 - e^{t_o b})}{e^{t_o a} - e^{t_o b}}, \quad E(T_i) \neq 0 = \frac{-ab}{\sigma^2_{T_i}}, \quad E(T_i) = 0$$

B. BSPD Under a Bivariate Lehmann Alternative

Let the dependence exist under the hypothesis H_o and alternative H_1. Assume the signal (represented by δ as the SNR) shifts the two-dimensional space according to the relationship[4]

$$H_o : F(x,y)$$

$$H_1 : G(x,y) = F^{1+\delta}(x,y) \approx F(x,y) + \delta F(x,y) \cdot \ell n\, F(x,y) \leq F(x,y)$$

The quantiles are assumed to be estimated under the univariate distribution $F(x)$. However, now the scores, b_{kj}, will reflect the bivariate dependence.

Define

$$P_{okj} = F(a_k, a_j) + F(a_{k-1}, a_{j-1}) - F(a_{k-1}, a_j) - F(a_k, a_{j-1})$$

as the probability of a sample z_i falling into cell (k,j). Then, the probability of falling into cell (k,j) under the selected alternative (H_{11}) and the alternative (H_1) are given by, respectively,

$$P_{1kj} \approx P_{okj} + \delta_1 A_{kj}$$

and

$$P_{kj} \approx P_{okj} + \delta A_{kj}$$

where δ and δ_1 have the same meaning as in the one-dimensional sequential partition detector discussed elsewhere in this report.

$$A_{kj} = F(a_k, a_j)\, \ell n\, F(a_k, a_j) + F(a_{k-1}, a_{j-1})\, \ell n\, F(a_{k-1}, a_{j-1})$$
$$- F(a_{k-1}, a_j)\, \ell n\, F(a_{k-1}, a_j) - F(a_k, a_{j-1})\, \ell n\, F(a_k, a_{j-1})$$

The mean and variance of T_i can be shown, straightforwardly, to be

$$E(T_i) = (\delta_1 \delta - \tfrac{1}{2}\delta_1^2) \sum_k^m \sum_j^m A_{kj}^2 / P_{okj} \tag{4}$$

and

$$\sigma_{T_i}^2 = \delta_1^2 \sum_k^m \sum_j^m A_{kj}^2 / P_{okj}$$

Then, the solution of $\phi(t) = 1$, $t \neq 0$, becomes,

$$\lim_{n\to\infty} t = t_o = -2\frac{E(T_i)}{\sigma^2_{T_i}} = (1 - 2\delta/\delta_1) = T(\delta) \tag{5}$$

This is an interesting result; $t(\delta)$ is free from any bivariate dependence. The thresholds remain constant along with α and β, which is a distinct advantage.

Instead of the thresholds, the scores, b_{kj}, will have to be adjusted to compensate for the amount of dependence introduced in any particular environment. Even though the BSPD is restricted to a sampling rate of $2/\Delta'$, where Δ' is the effective bandwidth, the form and amount of dependence need not be known per se. This means BSPD maintains constant α, β under unknown dependent noise conditions as long as the bivariate composition is maintained.

C. BSPD Under a Bivariate Shift Alternative

Assume the same conditions hold under a shift alternative as in B. Then,[4]

$$H_o : F(x,y)$$

$$H_1 : F(x,y) = F(x - \Delta, y - \Delta) \approx F(x,y) - \Delta\left(\frac{\partial}{\partial x} + \frac{\partial}{\partial y}\right) F(x,y)$$

$$H_{11} : G(x,y) \approx F(x,y) - \Delta_1\left(\frac{\partial}{\partial x} + \frac{\partial}{\partial y}\right) F(x,y)$$

where the quantiles again will be estimated under the univariate distribution $F(x)$.

Proceeding as before,

$$P_{okj} = F(a_k, a_j) + F(a_{k-1}, a_{j-1}) - F(a_{k-1}, a_j) - F(a_k, a_{j-1})$$

$$P_{1kj} \approx P_{okj} + \Delta_1 B_{kj}$$

$$P_{kj} \approx P_{okj} + \Delta B_{kj}$$

where

$$B_{kj} = [F_x(a_{k-1}, a_j) + F_y(a_{k-1}, a_j) + F_x(a_k, a_{j-1}) + F_y(a_k, a_{j-1}) - F_x(a_k, a_j) - F_y(a_k, a_j) - F_x(a_{k-1}, a_{j-1}) - F_y(a_{k-1}, a_{j-1})]$$

and

$$F_x(a_k, a_j) = \frac{\partial}{\partial x} F(x,y)\Big|_{\substack{x = a_k \\ y = a_j}}$$

$$F_y(a_k, a_j) = \frac{\partial}{\partial y} F(x,y)\Big|_{\substack{x=a_k \\ y=a_j}}$$

The mean and variance of T_i can be written directly,

$$E(T_i) = (\Delta\Delta_1 - \tfrac{1}{2}\Delta_1^2) \sum_k^m \sum_j^m B_{kj}^2 / P_{okj}$$

$$\sigma_{T_i}^2 = \Delta_1^2 \sum_k^m \sum_j^m B_{kj}^2 / P_{okj}$$

The solution for $t = t_o \neq 0$ becomes

$$\underset{n\to\infty}{t} = t_o = -2\frac{E(T_i)}{\sigma_{T_i}^2} = (1 - 2\Delta/\Delta_1) = t(\Delta)$$

where the Δ and Δ_1 have the same meaning as in the one-dimensional sequential test discussed elsewhere in this report.

Joint Services Technical Advisory Committee
F44620-74-C-0056
Naval Underwater System Command

R. F. Dwyer and L. Kurz

REFERENCES

1. Y. C. Ching and L. Kurz, "Nonparametric Detectors Based on m-Interval Partitioning," IEEE Trans. on Information Theory, IT-18, No. 2, 251-257 (March 1972).
2. W. Feller, "An Introduction to Probability Theory and Its Applications," John Wiley, 3rd Edition (1968).
3. T. Ferguson, "Mathematical Statistics: A Decision Theoretic Approach," Academic Press (1967).
4. G. Bhattacharyya and R. Johnson, "A Layer Rank Test for Ordered Bivariate Alternatives," Ann. Math. Stat., 41, No. 1, 1296-1310 (1970).

RELATIVE PERFORMANCE OF SEQUENTIAL AND FIXED SAMPLE SIZE PARTITION DETECTORS

R.F. Dwyer and L. Kurz

The main purpose of this report is to extend the theory of nonparametric partition detectors to include sequential operation. To that end, the observation space is partitioned into m intervals based on the knowledge of only a finite number of quantiles of the noise distribution, which can be easily estimated using the techniques presented in References 1 and 2. The ease of implementation and robustness of the m-interval detectors make the sequential partition detectors (SPD) particularly attractive.

Specifically, a formulation for the SPD is given and then test statistics are constructed for the Lehmann and shift-in-the-mean alternatives. Efficiency expressions are derived and maximized by finding the optimal quantiles of the noise distribution. Asymptotic properties for fixed sample size m-interval partition detectors are derived, and measures of efficiency are introduced which compare the fixed sample size detector with the SPD. An important theorem is proven which shows that the optimum quantiles which maximize the efficiency for fixed sample size and sequential partition detectors are the same. The latter fact permits the use of Fisher's information as a bound on performance and assures the existence of a locally most powerful (LMP) SPD. It is shown that an advantage of the SPD over its fixed sample size detector counterpart is that it can be up to four times more efficient based on the partition asymptotic relative efficiency (PARE).

A. Formulation of a Sequential Partition Detector (SPD)

Following Ching and Kurz,[3] SPD will be formulated by partitioning the observation space into a finite number of regions based on the quantiles of the noise distribution.

Before discussing the SPD in detail, a few remarks are needed. It is well known[4,5,6,7,8] that a sequential test based on the likelihood ratio will have the optimum property at θ_0 (signal strength under the hypothesis, H_0) and θ_1 (signal strength under the alternative, H_1). By estimating the quantiles of a noise distribution under H_0 and specifying an alternative sensitive to the signal strength θ_1, a likelihood ratio statistic is formed which is invariant to changing distributions. Since only the quantiles of the distribution need to be known, the form of the sequential test remains the same. Therefore, the objective of this section is to formulate a procedure which will minimize the average sample number (ASN) for the SPD as a function of the number of quantiles estimated. Let $E[n/\theta]$ and $E_0[n/\theta]$ represent the ASN for the SPD and the optimum SPD, respectively. We seek $\min[E(n/\theta)] \geq \min[E_0(n/\theta)]$, where n is a random variable representing the number of samples, and as $\lim_{m\to\infty} [\min E(n/\theta)] \to \min[E_0(n/\theta)]$, where (m - 1) are the number of quantiles estimated.

Assume $x_1, x_2, \cdots, x_n$ is a sample of i.i.d.r.v. with a a.c.d.f. F(x) and let $\vec{a} = (a_0 = -\infty, a_1, \cdots, a_{m-1}, a_m = \infty)$ be the quantiles of F(x). Let n_k be the number of observations falling into the interval $(a_{k-1}, a_k]$. In addition, introduce

$$P_{\nu k} = F_\nu(a_k) - F_\nu(a_{k-1}), \quad \nu = 0,1$$

representing the probability of a sample, x_i, falling into the k-th interval under the hypothesis or alternative subject to

$$\sum_{k=1}^{m} P_{\nu k} = 1 \quad \text{and} \quad \sum_{k=1}^{m} n_k = n$$

The likelihood ratio statistic upon which the SPD is based become

$$\frac{P_1(\vec{n})}{P_2(\vec{n})} = \frac{\prod_{k=1}^{m} P_{1k}^{n_k}}{\prod_{k=1}^{m} P_{0k}^{n_k}} \tag{1}$$

It is easy to show that an equivalent test to the one given by Eq. (1) is of the form

$$T_n = \sum_{i=1}^{n} \sum_{k=1}^{m} b_k n_{ik} = \sum_{i=1}^{n} T_i \tag{2}$$

where

$$n_k = \sum_{i=1}^{n} n_{ik}$$

and

$$b_k = \ell n(P_{1k}/P_{0k})$$

are the scores.

Equation (2) represents the classical cumulative sum form of the SPD with stopping boundaries given by

$$b = \ell n\, B < T_n < \ell n\, A = a \tag{3}$$

where the test is terminated if $T_n \geq a$ or $T_n \leq b$ and another sample (or samples) is (are) accumulated if T_n fall between a and b.

B. Constructing SPD's Based on Selected Alternatives

The problem now is to find a suitable alternative relating P_k, P_{1k} and P_{0k} to a meaningful parameter, δ, which separates the hypothesis from the alternative. Two such classes of alternatives will be discussed here: the Lehmann alternative (a discussion of this class is given elsewhere in this report) and shift-in-the-mean alternative.

1. Lehmann Alternative

Lehmann proposed a one parameter family of nonparametric classes of alternatives

$$G(x) = F^{1+\delta}(x) \tag{4}$$

where $G(x) \leq F(x)$ for $\delta > 0$. For δ small

$$H_0 : F(x) \quad \text{and} \quad H_1 : G(x) \approx F(x) + \delta F(x) \ln F(x)$$

The parameter δ represents the unknown true separation between the hypothesis and alternative under the Lehmann class. The objective of the SPD is either to accept or reject the hypothesis or accumulate another sample under the unknown separation δ given a preselected (usually based on the minimax criterion) separation δ_1. In this case

$$P_k = P_{0k} + \delta A_k$$

$$b_k = \ln\left(1 + \delta_1 \frac{A_k}{P_{0k}}\right)$$

and

$$A_k = F(a_k) \ln F(a_k) - F(a_{k-1}) \ln F(a_{k-1}) \tag{5}$$

Following the procedure of Ref. 3

$$E[T_i] = \ln(1 + \delta_1 A_k/P_{0k})(P_{0k} + \delta A_k) \tag{6}$$

and

$$\sigma^2_{T_i} = [\ln(1 + \delta_1 A_k/P_{0k})]^2 (P_{0k} + \delta A_k) - [E(T_i)]^2 \tag{7}$$

which reduce for δ_1 and δ small to

$$E[T_i] = (\delta \delta_1 - \tfrac{1}{2} \delta_1^2) \sum_{k=1}^{m} A_k^2/P_{0k} \tag{8}$$

and

$$\sigma_{T_i}^2 = \delta_1^2 \sum_{k=1}^{m} A_k^2/P_{0k} \tag{9}$$

Following Wald[5] and Bahadur[7]

$$t_o(\delta) = 1 - 2\delta/\delta_1$$

and the performance measures - probability of terminating the test and ASN - are

$$P(\delta) = \frac{1 - e^{t_o(\delta)\, b}}{e^{a\, t_o(\delta)} - e^{b\, t_o(\delta)}} \tag{10}$$

and

$$E[n/\delta] = \frac{1}{(\delta\delta_1 - \frac{1}{2}\delta_1^2)} \left\{ \frac{1}{N_0 \sum_{k=1}^{m} A_k^2/P_{0k}} \cdot \right.$$

$$\left. \frac{b(e^{a\, t_o(\delta)} - 1) + a(1 - e^{b\, t_o(\delta)})}{e^{a\, t_o(\delta)} - e^{b\, t_o(\delta)}} \right\}, \quad \delta \neq \delta_1^2/2$$

$$= \frac{-ab}{\delta_1^2 N_0 \sum_{k=1}^{m} A_k^2/P_{0k}}, \quad \delta = \delta_1^2/2 = \delta_c \tag{11}$$

It is interesting to note that in Eqs. (10) and (11) δ and δ_1 do not depend on the quantiles. This means that $E[n/\delta]$ can be minimized by maximizing

$$\sum_{k=1}^{m} A_k^2/P_{0k} \cdot$$

2. Shift Alternative

In this case the alternative is related to the hypothesis by

$$G(x) = F(x - \Delta)$$

where Δ represents the shift or location parameter in the noise distribution, $F(x)$. If Δ and Δ_1 have similar meaning as δ and δ_1 for Lehmann alternatives, for small value of these parameters

$$E[T_i] = (\Delta\Delta_1 - \frac{1}{2}\Delta_1^2) \sum_{k=1}^{m} \frac{[f(a_{k-1}) - f(a_k)]^2}{F(a_k) - F(a_{k-1})} \tag{12}$$

and

$$\sigma_{T_i}^2 = \Delta_1^2 \sum_{k=1}^{m} \frac{[f(a_{k-1}) - f(a_k)]^2}{F(a_k) - F(a_{k-1})} \tag{13}$$

which are similar to Eqs. (8) and (9) except that the slope of F(x) or the p.d.f is involved. If one identifies

$$\sum_{k=1}^{m} A_k^2/P_{0k} = \sum_{k=1}^{m} \frac{[f(a_{k-1}) - f(a_k)]^2}{F(a_k) - F(a_{k-1})} \tag{14}$$

the optimization procedure for Lehmann alternatives may be followed.

C. Measures of Efficiency for Sequential and Fixed Sample Size Partition Detectors

One of the important results of comparison between SPD and fixed sample size partition detectors may be summarized in the form of a theorem.

Theorem 1: The optimum quantiles which maximize the asymptotic relative efficiency (ARE) for the SPD and the fixed sample size partition detector under Lehmann and shift alternatives are the same.

Proof: See Reference 9.

This is an important result which assures optimization (for some m) of both the SPD and fixed sample size partition detectors for the same test statistic under identical alternatives. Another interesting point is the efficacy of the SPD is bounded by Fisher's information the same as for the fixed sample size partition detectors.[3] Thus, a locally most powerful SPD exists based on quantiles or as $m \to \infty$, the SPD approaches the best among all locally optimum sequential tests for the same noise environment. Since the quantiles are estimated under ambient noise conditions, the SPD possesses the desirable property of being able to adapt to changing noise conditions and maintain its optimum quantities. Other sequential tests designed under known distributions may perform poorly in a changing noise environment.

The disadvantage of the Pitman-Noether ARE approach as a performance measure is that it does not show the dependence on α and β over a range of signal strength, θ. Bahadur and Hodges-Lehmann efficiencies alleviate this difficulty somewhat.[3]

However, a better approach is needed when comparing sequential and nonsequential detectors.

Consider a detector of the form

$$T_n = \sum_{i=1}^{N} T_i \tag{15}$$

as in Equation (2). If α and β are small then

$$N = \frac{[\sqrt{-2\ell n\,\alpha} + \sqrt{-2\ell n\,\beta}}{\theta_1^2 \sum_{k=1}^{m} A_k^2/P_{0k}} \tag{16}$$

The ratio of $N/E(n/\theta)$ as $\alpha, \beta \to 0$ will be defined as partition asymptotic relative efficiency (PARE)

$$\text{PARE} = \lim_{\alpha=\beta\to 0} \frac{N}{E(n/\theta)} = \lim_{\alpha=\beta\to 0} \frac{-\frac{1}{2}[\sqrt{-2\ell n\,\alpha} + \sqrt{-2\ell n\,\beta}]^2 t_o(\theta)}{\frac{b(e^{a t_o(\theta)} - 1) + a(1 - e^{b t_o(\theta)})}{e^{a t_o(\theta)} - e^{b t_o(\theta)}}} \tag{17}$$

The PARE has similarities with the Pitman-Noether ARE in that both let the respective sample sizes $\to\infty$ but in different ways. The Pitman-Noether ARE lets $H_1 \to H_0$ while fixing α and β and PARE lets $\alpha, \beta \to 0$ for some θ. However, both are independent of α and β asymptotically.

Instead of letting $\alpha = \beta \to 0$, it is convenient to let $a = |b| \to \infty$ thus obtaining

$$\text{PARE} = \begin{cases} 4 & t_o(\theta) = 1 \\ 4 & t_o(\theta) = -1 \\ \to\infty & t_o(\theta) \to -\infty \\ 0 & t_o(\theta) = t_o(\theta_c) \end{cases} \tag{18}$$

At $\theta = \theta_0$ or $\theta = \theta_1$ the SPD is four times as efficient in terms of PARE as its fixed sample size counterpart. For $\theta_0 < \theta < \theta_1$, the SPD loses its advantage over the fixed sample size detector.

In practical situation one wishes to have finite truncation time. In the latter situation procedures suggested by Bussgang and Middleton,[10] Weiss,[11] Anderson,[12] Kiefer and Weiss[13] and Bussgang and Marcus[14] are applicable.

Joint Services Technical Advisory Committee
F44620-74-C-0056
Naval Underwater System Command

R.F. Dwyer and L. Kurz

REFERENCES

1. P. Kersten and L. Kurz, "Robustized Vector Robbins-Monro Algorithm with Applications to M-Interval Detection," Information Sciences, to appear.
2. P. Kersten and L. Kurz, "Bivariate m-Interval Classifiers with Application to Edge Detection," Information and Control, to appear.
3. Y.C. Ching and L. Kurz, "Nonparametric Detectors Based on M-Interval Partitioning," IEEE Trans. on Information Theory, IT-18, No. 2, 251-257 (March 1972).
4. A. Wald and J. Wolfowitz, "Optimum Character of the Optimum Sequential Probability Ratio Test," Ann. Math. Stat., 19, No. 3 (September 1948).
5. A. Wald, "Sequential Analysis," (New York: John Wiley, 1947).
6. T. Ferguson, "Mathematical Statistics: A Decision Theoretic Approach," Academic Press (1967).
7. R. Bahadur, "A Note on the Fundamental Identity of Sequential Analysis," Ann. Math. Stat., 29, 534-543 (1958).
8. D. Blackwell and M. Girshick, "Theory of Games and Statistical Decisions," John Wiley (1954).
9. R.F. Dwyer, "Sequential Partition Detectors with Applications," Ph.D. Dissertation, Polytech. Inst. of New York (June 1976).
10. J. Bussgang and D. Middleton, "Optimum Sequential Detection of Signals in Noise," IRE Trans. on Information Theory, IT-1 (December 1955).
11. L. Weiss, "Testing One Simple Hypothesis Against Another," Ann. Math. Stat., 24 24 (1953).
12. R. Anderson, "A Modification of the Sequential Probability Ratio Test to Reduce the Sample Size," Ann. Math. Stat., 31 (1960).
13. J. Kiefer and L. Weiss, "Some Properties of the Generalized Sequential Probability Ratio Test," Ann. Math. Stat., 28 (1957).
14. J. Bussgang and M. Marcus, "Truncated Sequential Hypothesis Tests," IEEE Trans. on Information Theory, IT-13, No. 3 (July 1967).

A FLEXIBLE MODEL FOR IMPULSIVE NOISE

L. Kurz

In most data transmission systems the noise distribution is appropriately modelled by the mixture distribution consisting of gaussian and impulsive noise. Though the available data on statistical behavior of impulsive noise is limited, it is generally recognized that the impulsive noise may be represented by clusters of impulses with near-Poisson distribution of occurrence and amplitude distribution with large variance. In view of this limited knowledge, representations and procedures are suggested here which permit to represent the impulsive noise and mixture distribution by a flexible model both in terms of occurrence and amplitude distributions. In addition, a recursive procedure for estimating the mixture parameter is also given.

A. A Model for Amplitude Distribution of Impulsive Noise

In this section, a one-dimensional form of a representation theory, which results in good fit with small number of terms, will be derived. The representation theory is based on the concepts leading to the Gram-Charlier type A series and is similar to the development of Ref. 1 for dispersive channels.

Consider a generalized Fourier series

$$\hat{p}(x) = \sum_{i=1}^{\infty} c_i \phi_i(x)\, p(x) \tag{1}$$

where $p(x)$ is some property preselected reference p.d.f. and $\hat{p}(x)$ represents an estimator of the actual p.d.f. The choice of the orthonormal set $\{\phi_i(x)\}$ corresponds to the choice of $p(x)$ and is based on the approach given by Szego.[2]

As suggested by Algazi and Lerner,[3] a good choice for $p(x)$ is

$$p(x) = a\, e^{-b|x|^{m/n}} \tag{2}$$

where

$$.1 < \frac{m}{n} < 1$$

Following Szego[2] the orthoncrmal polynomials $\{\phi_i(x)\}$ are generated from the moments, m_n, of the reference distribution by

$$\phi_n(x) = (D_{n-1}D_n)^{-\frac{1}{2}} \begin{vmatrix} m_0 & m_1 & m_2 & \cdots & m_n \\ m_1 & m_2 & m_3 & \cdots & m_{n+1} \\ \vdots & & & & \\ m_{n-1} & m_n & m_{n+1} & \cdots & m_{2n-1} \\ 1 & x & x^2 & \cdots & x^n \end{vmatrix} \tag{3}$$

where for $n > 0$

$$D_n = \left[m_{\nu+\mu}\right]_{\nu,\mu} = 0, 1, 2, \cdots, n$$

The first four orthogonal polynomials are then

$$\phi_0(x) = 1$$

$$\phi_1(x) = \frac{1}{\sqrt{m_2}}\, x$$

$$\phi_2(x) = \frac{1}{\sqrt{m_4 - m_2^2}}\, x^2 - \frac{m_2}{\sqrt{m_4 - m_2^2}}$$

$$\phi_3(x) = \frac{m_2}{\sqrt{m_2^2 m_6 - m_2 m_4^2}}\, x^3 - \frac{m_4}{\sqrt{m_2^2 m_6 - m_2 m_4^2}}\, x$$

$$\phi_4(x) = \frac{m_4 - m_2^2}{c}\left\{x^4 + \left[\frac{m_4 m_2 - m_6}{m_4 - m_2^2}\right]x^2 - \left[m_4 + m_2\,\frac{m_4 m_2 - m_6}{m_4 - m_2^2}\right]\right\}$$

where

$$c = m_8(m_4 - m_2^2) + 2m_6(m_4 m_2 - m_6) + m_4\left[\frac{m_4 m_2 - m_6}{m_4 - m_2^2}\right]$$

$$- \left[\frac{m_4 m_2 - m_6}{m_4 - m_2^2}\right] 2m_2(m_4^2 - m_2 m_6) + \frac{[m_4^2 - m_2 m_6]^2}{m_4 - m_2^2} - 2(m_4^2 - m_2 m_6)m_4$$

If the unknown p.d.f. $\hat{p}(x)$ is of the form

$$\hat{p}(x) = a_1 \exp - b_1 |x|^{1/m} \tag{4}$$

and the reference p.d.f.

$$p(x) = a_2 \exp - b_2 |x|^{1/n} \tag{5}$$

with

$$\hat{p}(x) = \sum_{i=1}^{N} c_i \phi_i(x)\, p(x) \tag{6}$$

then it can be shown that for the approximation to be valid

$$n > m\left[\frac{N+1}{2N+1}\right] \tag{7}$$

Typically, N = 4 gives the relationship between n and m for an acceptable approximation.

Consider the mixture distribution noise model with p.d.f.

$$p_m(x) = (1 - q)\, p_g(x) + q\, p_i(x) \tag{8}$$

where $p_m(\cdot)$, $p_g(\cdot)$ and $p_i(\cdot)$ are the amplitude p.d.f. of mixture, gaussian and impulsive noise components, respectively.

If $p_m(x)$ is of the form

$$p_m(x) = a_m \exp - b_m |x|^{1/\nu}$$

and

$$p_i(x) = a_i \exp - b_i |x|^{1/n}$$

it is easy to show that, for a given q and n, the value of ν such that

$$(s_{\nu-1})! > q\,[(s_{n-1})!] \tag{9}$$

For instance, for N = 4 and n = 1, for an actual q = .03 and $\nu = 1/2$, the bound Eq. (9) yields $\hat{q} = .053$.

The one-dimensional representation may be extended to two or more dimensions following the procedure of Reference 4.

Masry[5] points out that the two-dimensional expansions of this type are not unique nor need they converge to the actual joint p.d.f. This convergence is not needed in practical situations. What is needed is that a finite number of terms in the series represent the desired p.d.f. adequately for a given application. The latter goal is almost always achieved.

B. Representation of the Time of Occurrence of Impulsive Noise and Estimation of the Mixing Parameter

Following the philosophy of Section A, the frequency of occurrence of the noise impulses may be described by a Gram-Charlier type B series.[6-8] Using the Poisson distribution as the reference measure, the frequency distribution of the impulsive noise may be represented by

$$P(x=k) = f_o(k;\lambda) \sum_{i=0}^{\infty} a_i C_i(k;\lambda) \tag{10}$$

where

$$f_o(k;\lambda) = \lambda^k e^{-\lambda}/k!$$

$$C_i(k;\lambda) = \frac{i!}{\lambda^i} L_i^{(k-i)}(\lambda)$$

and

$$L_i^{(\nu)}(\lambda) = \frac{e^{\lambda}}{\lambda^{\nu} i!} \frac{d^i}{d\lambda^i}\left[e^{-\lambda}\lambda^{i+\nu}\right]$$

If a finite number of terms of Eq. (10), say K, represent adequately the frequency of occurrence of the impulses, all the parameters necessary for this representation may be obtained by determining or estimating the K+1 moments of the distribution following a procedure similar to the one presented in Section A.

In general, the problem of finding a suitable estimator of the mixing parameter, q, based on independent signal-free samples, is a problem of optimizing a suitable functional.

It is easy to show that for the model of Eq. (8), using the minimum integral-square error criterion, the problem reduces to minimization of

$$I(q) = \int_{-\infty}^{\infty} [p_m(x) - (1-q)\,p_g(x) - q\,p_i(x)]^2\,dx \tag{11}$$

which following the usual procedures of the calculus of variations, yields

$$\hat{q} = -\frac{\int_{-\infty}^{\infty} p_m(x)\,[p_g(x) - p_i(x)]\,dx + c_2}{c_1} \tag{12}$$

where

$$c_1 = \int_{-\infty}^{\infty} [f_g(x) - f_i(x)]^2 dx$$

and

$$c_2 = \int_{-\infty}^{\infty} [f_g(x)\, f_i(x) - f_g^2(x)]\, dx$$

Introducing a random function $Z_{P_m}(\xi, x)$ defined as

$$Z_{P_m}(\xi, x) = \begin{cases} 1 & \text{if } \xi \leq x \\ 0 & \text{otherwise} \end{cases}$$

it can be seen that

$$E_\xi [Z_{P_m}(\xi, x)] = P_m(x)$$

where $P_m(x)$ is the c.d.f. of the mixture distribution. Thus, the regression equation

$$E_\xi [y(q, x)/q] = E_\xi \left[\int_{-\infty}^{\infty} \frac{d Z_{P_m}(\xi, x)}{d\xi} [p_g(x) - p_i(x)]\, dx + q C_1 + C_2 \right]$$

$$= [p_g(x) - p_i(x)] + q C_1 + C_2 \qquad (13)$$

which is equivalent to finding the minimum of Equation (11).

The problem of finding $\hat{q}$ of Eq. (12) is, therefore, equivalent to the problem of finding the zero of regression Equation (13). Following the Robbins-Monro procedure,[9-13] an iterative solution for $\hat{q}$ is obtained from observations $\{y_i\}$ in the form

$$\hat{q}_{i+1} = \hat{q}_i - a_i y(\hat{q}_i, y_i) \qquad (14)$$

where $a_i = a/i$ and $a > 0$.

Noting that $q \in [0, 1]$, the random function $y(\hat{q}_i, x)$ is redefined as in[11]

$$y(\hat{q}_i, x) = \begin{cases} -C_q & \text{if } \hat{q}_i < 0 \\ [f_g(x) - f_i(x)] + \hat{q}_i C_1 + C_2 & \text{if } 0 \leq \hat{q}_i \leq 1 \\ +C_q & \text{if } q_i > 1 \end{cases} \qquad (15)$$

where C_q is a suitable positive constant. It is readily shown that with the choice of $y(\hat{q}_i, x)$ given by Eq. (15), all conditions for the Robbins-Monro algorithm are satisfied. Thus, the recursive estimator of Eq. (14) converges to the optimum estimator in mean

square sense and with probability one. The usual acceleration procedures are also applicable.

In practical situations, the $f_i(x)$ of Eq. (15) is not known. In the latter case, $f_i(x)$ may be replaced by $\hat{f}_i(x)$ as found using the method of Section A.

Joint Services Technical Advisory Committee
F44620-74-C-0056

L. Kurz

REFERENCES

1. D.R. Klose and L. Kurz, "A New Representation Theory and Detection Procedures for a Class of Non-Gaussian Channels," IEEE Trans. on Com. Tech., COM-17, No. 2, 225-235 (April 1969).
2. G. Szego, "Orthogonal Polynomials," American Mathematical Society Colloquium Publications, XXIII (1939).
3. V.R. Algazi and R.M. Lerner, "Binary Detection in White Non-Gaussian Noise," M.I.T. Lincoln Laboratory, Technical Report DS-2138.
4. D.R. Klose and L. Kurz, "A Multivariable Representation Theory and Detection Procedures for a Class of Nongaussian Stochastic Processes," Proc. 1968 Princeton Conference on Information Sciences, 175 (March 1968).
5. E. Masry, "On the Series Representation of Bivariate Probability Density Functions," Proc. IEEE, 57, No. 10, 1773-1774 (October 1969).
6. A. Fisher, "The Mathematical Theory of Probability," 1, McMillan Co. (1936).
7. M. Kendal and A. Stuart, "The Advanced Theory of Statistics," 1, Hafner Publishing Co. (1958).
8. H. Levenbach, "Approximate Zero-Crossing Distributions Using the Type B Gram-Charlier Series," IEEE Trans. on Info. Theory, 11, 507-512 (October 1965).
9. N.V. Loginov, "Methods of Stochastic Approximation," Automatika i Telemekhamika, 27, No. 4, 185-204 (April 1966).
10. H. Robbins and S. Monro, "A Stochastic Approximation," Ann. Math. Stat., 22, 400-407 (1951).
11. A. Dworetzky, "On Stochastic Approximations," Proc. Third Berkeley Symp. on Math. Stat. and Prob., 1, Univ. of Calif. Press, 39-55 (1956).
12. J.L. Hodges and E.L. Lehmann, "Two Approximations of the Robbins-Monro Process," Proc. Third Berkeley Symp. on Math. Stat. and Prob., 1, Univ. of Calif. Press, 95-104 (1956).
13. L. Schmetterer, "Stochastic Approximations," Proc. Fourth Berkeley Symp. on Math. Stat. and Prob., 1, Univ. of Calif. Press, 587-609 (1961).

SUBOPTIMUM DETECTION IN NONSTATIONARY GAUSSIAN NOISE

L. Kurz

The problem of detection and estimation in nonstationary gaussian noise has been considered by many researchers. The approaches used to solve the problem lead to complicated mathematical formulations. Some of the difficulties become clear if one considers just one such approach presented elsewhere in this report. The purpose of this paper is to develop a suboptimum procedure to solve the problem. In particular, an approach originally suggested by Tukey[1] and modified by Priestley[2,3] is used to generate a set of observables which permit a reasonable solution to the problem. In addition, the techniques suggested here help one understand better the physical phenomena associated with the actual detection and estimation process.

A. Generation of the Observables

As the first step in the development of a meaningful procedure for comparing two different gaussian stochastic processes, an evolutionary spectral estimator is generated from the observed waveshapes. This procedure, originally suggested by Tukey,[1] involves a process of complex demodulation. The actual complex demodulation is performed in two steps:

1) The observed sample function of the gaussian stochastic process, x(t), is demodulated by multiplying it by $\exp(-j\omega_0 t)$, where ω_0 is the demodulation radian frequency.

2) The demodulated version of the observed sample function is smoothed by a low-pass discrete linear filter, with impulse response h_k producing

$$V(t) = \sum_{k=1}^{K} h_k X(t-k) \exp[-j\omega_0 (t-k)] \tag{1}$$

where V(t) is the complex demodulate.

Priestley[2] improved on the Tukey's procedure by introducing a second low-pass filter to the modulus of the complex demodulate. The evolutionary spectral estimator at time t and radian frequency ω_0 is then

$$f_{t,x}(\omega_0) = \sum_{k=1}^{K} w_k V(t-k)\, V^*(t-k) \tag{2}$$

where w_k is the second low-pass smoothing filter.

The evolutionary cross-spectrum is estimated as follows:

1) For each sequence $X(t_1), X(t_2), \cdots, X(t_n)$ calculate the complex demodulate $V_{x_i}(t)$ using Equation (1)

2) Form an n x n matrix whose elements are the products $V_{x_i}(t) V^*_{x_k}(t)$, where $i, k = 1, 2, \cdots, n$

3) Smooth the matrix of cross-products with a low-pass filter w_k.

Using these steps, the evolutionary cross-spectral density matrix is generated with terms of the form given by Equation (2). The evolutionary cross-spectral density matrix may be treated as an estimate of a multivariate normal covariance matrix.

The two-stage filtering process suggested by Priestly may be considered as corresponding to a two-dimensional filter. The first stage represents the same operation as encountered in AM receivers. This operation singles out the neighborhood of one frequency for analysis. Once the time-varying power in the neighborhood of this frequency is isolated, the second filter is used for smoothing purposes. The choice of the parameters of the two-dimensional filter is presented in Ref. 3 and will not be given here.

Consider a univariate evolutionary process with sample function

$$X(t) = c(t)\, \psi(B)\, \varepsilon(t) \tag{3}$$

where

$$c(t) = \begin{cases} \alpha & \text{for } t \varepsilon T_i \text{ for all odd } i \\ \beta & \text{for } t \varepsilon T_i \text{ for all even } i \end{cases}$$

and α, β are real constants. In Eq. (3), $\psi(B)$ is a rational function of B and $\varepsilon(t)$ is a sample function of a white noise process. The product $\psi(B)\varepsilon(t)$ may be interpreted as the stationary part of X(t) which is modulated by c(t). The design of the evolutionary process estimator requires knowledge of the change in c(t) with respect to time and the change in the stationary part $\psi(B)\varepsilon(t)$ with respect to frequency. These two characteristics of X(t) are measured by a time domain bandwidth, B_c, and the more traditional frequency domain bandwidth, B_ω, respectively. The time domain bandwidth may be considered to be a measure of the maximum time interval over which X(t) may be treated as approximately stationary.

The T_i are selected as interarrival times of a Poisson process with a mean λ. The Poisson assumption implies that the only available information on the occurrence of the next event in the knowledge of the mean time of such occurrences. For further details of the design procedure one should consult Reference 3.

B. The Detection Problem

The detection problem of comparing two evolutionary processes is posed as a

problem of hypothesis testing. In particular, given a sample function of a process X(t) known to be generated by a linear transformation of one of the two processes Y(t) and Z(t) for $t \varepsilon T_i$, then X(t) ~ Y(t) is the hypothesis, H_0, and X(t) ~ Z(t) is the alternative, H_1. As the test statistic to test the hypothesis one introduces

$$T(x) = r_{xy/z} - r_{xz/y} \tag{4}$$

where the r's are the appropriate partial coherence coefficients. These coefficients are calculated using Priestley's procedure. The estimates of r's following this procedure are then

$$q(t) = \hat{r}_{xy/z} - \hat{r}_{xz/y} = T(t) + \xi(t) \tag{5}$$

where $\xi(t)$ is a sample function of colored gaussian noise. The process, $\xi(t)$, is related to the white gaussian process, $\varepsilon(t)$, by

$$N(B)\,\xi(t) = D(B)\,\varepsilon(t) \tag{6}$$

where N(B) and D(B) are polynomials in B. Equation (6) represents the autoregressive moving average (ARMA). The actual hypothesis testing is then

$$(\underline{T}^{(1)} - \underline{T}^{(0)})^t\, \Lambda^{-1} q > \Lambda\,, \quad \Lambda_0 > 0 \tag{7}$$

implies that H_0 is rejected and the threshold, Λ_0, is selected in the usual manner. In Eq. (7)

$$H_0 : T = T^{(0)} \text{ for all } t \varepsilon T_i$$

$$H_1 : T = T^{(1)} \text{ for all } t \varepsilon T_i$$

$$q^t = (q_1, q_2, \cdots, q_{T_i})\,, \quad \xi^t = (\xi_1, \xi_2, \cdots, \xi_{T_i})$$

$$\underline{T}^t = (T_1, T_2, \cdots, T_{T_i}) \quad \text{and} \quad \Lambda = E[\xi\,\xi^t]$$

The problem of inverting Λ may be solved by the numerous available computer-oriented techniques.

Joint Services Technical Advisory Committee
F44620-74-C-0056

L. Kurz

REFERENCES

1. J.W. Tukey, "Discussion: Emphasizing the Connection Between Analysis of Variance and Spectrum Analysis," Technometrics, 3, 191-228 (1961).
2. M.B. Priestley, "Evolutionary Spectra and Non-Stationary Processes," J. of the Royal Statistical Society, series B, 7, No. 2, 204-237 (1965).
3. M.B. Priestley, "Design Relations for Non-Stationary Processes," J. of the Royal Statistical Society, series B, 28, No. 1, 228-240 (1966).

ON THE PARTITIONING REPRESENTATION FOR FINITE SETS AND ITS APPLICATION

T. Lee and E.J. Smith

A set of rules for representing the partitions on a finite set is developed and algorithms for computing the least upper and greatest lower bounds are given. Several applications including methods for determining the set of partitions having the substitution property, the set of partition pairs, and the minimal reduced machine for a given finite-state sequential machine are illustrated. The approach is also extended to find the subsemigroups of a given semigroup.

The concepts of equivalence relation and partitions are widely used in dealing with sequential machines and yield fruitful results in decomposition theory;[7] however, the computation procedure is complicated and troublesome, especially in the case of incompletely specified machines. The purpose of the present work is to develop a general approach for computing the desired partitions with the aid of a digital computer. In the following summary, proofs of theorems and lemmas are omitted for brevity.[12]

Hutchinson[1] has given a set of rules for representing the partitions on a finite set of n elements by an n-tuple integer array and an algorithm for generating these partitions; a "loopless" algorithm for generating the partitions is also developed by Ehrlich.[2,3] We will define an alternate set of rules for the representation of partitions which preserves the partial ordering. The conversion between the two representations is straightforward.

A partition π on a finite set X is a collection of mutually disjoint non-empty subset of X whose union is X. The members of π are called blocks or equivalence classes and the existence of two elements a, b of X in the same block of π is denoted by $a \equiv b(\pi)$. The n elements of a given finite set are numbered $1, 2, \ldots, n$. Without loss of generality, let X be the set of integers $\{1, 2, \ldots, n\}$ and let the set of all partitions on X be $\Pi(X)$. The representation of a partition $\pi = \{\overline{B}_1, \overline{B}_2, \ldots, \overline{B}_m\}$ on the set X is an n-tuple integer array $(a_1, a_2, \ldots, a_n)$ such that for $i \in \overline{B}_k$, $a_i = \min \overline{B}_k$, where $\min \overline{B}_k$ is the smallest element in $\overline{B}_k$.

Suppose n = 7, then the partition

$$\pi = \{\overline{1, 3, 6} \ , \ \overline{2, 5} \ , \ \overline{4, 7}\}$$

is represented by the 7-tuple array (1, 2, 1, 4, 2, 1, 4). We may consider the n-tuple $(a_1, a_2, \ldots, a_n)$ as an ordered set or image corresponding to a mapping from the set X into itself, $f: X \to X$, such that $(a_1, a_2, \ldots, a_n) = (f(1), f(2), \ldots, f(x))$. The question arises as to whether the mapping $f: X \to X$ is a valid representation of the partition on the set X. A mapping $f: X \to X$ represents a partition on X satisfying the following criteria:

(1) $f(i) \leq i$ for $1 \leq i \leq n$ (contraction)

(2) $f \cdot f(i) = f(i)$ for $1 \leq i \leq n$ (idempotency)

Where $\cdot$ denotes the functional composition, i.e., $f \cdot g(i) = f(g(i))$.

Let $F(X)$ be the set of all mappings from X into itself which satisfy the criteria of contraction (1) and idempotency (2).

Definition: Let $\Phi : F(X) \rightarrow \Pi(X)$ for $f \in F(X)$ and $\Phi(f) = \pi \in \Pi(X)$ iff

$f(i) = f(j) \leftrightarrow i \equiv j(\pi)$, $\forall$ i, j $\in X$

Lemma 1: Given $\pi \equiv \{\overline{B}_1, \overline{B}_2, \ldots, \overline{B}_m\} \in \Pi(X)$ and $f \in F(X)$ such that $\Phi(f) = \pi$, then $f(i) = \min \overline{B}_k$ for $1 \leq i \leq n$ and $i \in \overline{B}_k$.

Lemma 2: $\Phi : F(X) \rightarrow \Pi(X)$ is one-to-one and onto.

Definition: For $f_1, f_2 \in F(X)$, if $f_1(i) = f_1(j)$ implies $f_2(i) = f_2(j)$, $\forall$ i, j $\in X$, then $f_1 \leq f_2$. The set $F(X)$ with the partial ordering $\leq$ is a partially ordered set (POS).

Theorem 1: Given $f_1, f_2 \in F(X)$ then $f_1 < f_2$ iff $f_2 \cdot f_1 = f_2$.

Corollary 1: If $f_1 \leq f_2$, then $f_2(i) \leq f_1(i)$ for all $1 \leq i \leq n$.

Definition: A mapping θ: L $\rightarrow$ M from a POS L to a POS M is called an order-preserving mapping iff, $\forall$ a, b $\in$ L, $a \leq b$ implies $\theta(a) \leq \theta(b)$.

Lemma 3: The mapping $\Phi : F(X) \rightarrow \Pi(X)$ is an order-preserving mapping.

The least upper bound of $f_1, f_2 \in F(X)$ is $f = \text{lub}(f_1, f_2) = f_1 \vee f_2$ such that $f \geq f_1$, $f \geq f_2$, and for any $g \in F(X)$, $g \geq f_1$, $g \geq f_2$ implies $g \geq f$. The definition for greatest lower bound $f = \text{glb}(f_1, f_2) = f_1 \wedge f_2$ is similar. The following two theorems show the existence of lub and glb, and lead to the algorithms for computing them.

Theorem 2: For $f_1, f_2 \in F(X)$ let

$h_0(i) = f_2 \cdot f_1(i)$

$h_1(i) = h_0 \cdot f_2(i)$

$h_2(i) = h_1 \cdot f_1(i)$

$\vdots$

$h_k(i) = h_{k-1} \cdot f_s(i)$ for all $1 \leq i \leq n$

where

$$s = \begin{cases} 1 \text{ if } k \text{ is even} \\ 2 \text{ if } k \text{ is odd} \end{cases}$$

and k is the smallest integer such that $h_k = h_{n+1}$; then $f = h_k = \text{lub}(f_1, f_2)$.

Algorithm: lub

For $f_1, f_2 \in F(X)$ and $f = \text{lub}(f_1, f_2)$, let $A(i) = f_1(i)$, $B(i) = f_2(i)$ and $X(i) = F(i)$ for $1 \leq i \leq n$; the algorithm follows.

1. Read A(i), B(i) for $1 \leq i \leq n$.
2. Set X(i) = A(i) and Y(i) = B(i) for $1 \leq i \leq n$.
3. Set k = -1.
4. If X(i) = Y(i) for $1 \leq i \leq n$, then go to step 6; otherwise set k = k+1.
5. If k is even then set X(i) = Y(A(i)) for $1 \leq i \leq n$ and go to step 4; otherwise set Y(i) = X(B(i)) for $1 \leq i \leq n$ and go to step 4.
6. Record X(i) for $1 \leq i \leq n$.

Theorem 3: For $f_1, f_2 \in F(X)$. Let f(1) = 1 and for $i \geq 2$, define f(i) recursively,

$$f(i) = \begin{cases} f(k) & \text{if } f_1(i) = f_1(k) \text{ and } f_2(i) = f_2(k) \text{ for some } k,\ 1 \leq k < i. \\ i & \text{otherwise} \end{cases}$$

then $f = \text{glb}(f_1, f_2)$.

Algorithm: glb

1. Read A(i), B(i) for $1 \leq i \leq n$.
2. Set i = 2 and X(1) = 1.
3. Set k = 1.
4. If A(i) = A(k) and B(i) = B(k), then set x(i) = x(k), go to step 6; otherwise set k = k+1.
5. If k = i, then set x(i) = i and go to step 6; otherwise go to step 4.
6. Set i = i+1.
7. If $i \leq n$ then go to step 3; otherwise record x(i) for $1 \leq i \leq n$.

The triple $\{F(X), \vee, \wedge\}$ is a lattice and, moreover, the mapping $\Phi : F(X) \to \Pi(X)$ is a morphism of lattices.[4]

Theorem 4: For $f_1, f_2 \in F(X)$ and $f = \text{lub}(f_1, f_2) = f_1 \vee f_2$, $g = \text{glb}(f_1, f_2) = f_1 \wedge f_2$, then $\Phi(f) = \Phi(f_1) \vee \Phi(f_2)$ and $\Phi(g) = \Phi(f_1) \wedge \Phi(f_2)$.

A. Applications

Partitions with the substitution property are essential tools in the study of sequential machines. In this section we will discuss several examples of algorithms for generating partitions having various properties. The approach used in obtaining these algorithms is unified so that the results can be extended to that of finding the quotient structures of a given semigroup as well as of other finite algebraic system.

A machine (Mealy) is a 5-tuple

$\{S, X, Z, \delta, \lambda\}$

where

(1) $S = \{1, 2, \ldots, n\}$ is the set of states of size n
(2) $X = \{1, 2, \ldots, p\}$ is the set of input alphabets of size p
(3) $Z = \{1, 2, \ldots, m\}$ is the set of output alphabets of size m
(4) $\delta : S \times X \rightarrow S$ is the next state function
(5) $\lambda : S \times X \rightarrow Z$ is the output function.

We now consider the case of deterministic machines. X(i), Y(i) for $1 \leq i \leq n$ are n-tuple variable arrays, NX(i,k), Z(i,k) for $1 \leq i \leq n$, $1 \leq k \leq p$ are two-dimensional arrays, and A(i,j) for $1 \leq i \leq n$, $1 \leq j \leq n$, B(i,j,k), C(i,j,k) for $1 \leq i \leq n$, $1 \leq j \leq n$, $1 \leq k \leq m$ are Boolean arrays.

1. The Algorithm for Generating All Partitions with the Substitution Property

Definition: A partition π on the set of states of machine M is said to have the substitution property iff, $\forall\, k \in X$, $\forall\, i, j \in S$,

$$(i \equiv j(\pi)) \supset (\delta(i,k) \equiv \delta(j,k)(\pi)).$$

Suppose $f \in F(x)$ and $\Phi(f) \equiv \pi$, then we say f has the substitution property iff

$$(\forall k \in X)(\forall i, j \in S),\ (f(i) = f(j)) \supset (f(\delta(i,k)) = f(\delta(j,k))) \tag{1}$$

We may call Eq. (1) the logic equation, that is to say any $f \in F(S)$ has the substitution property if and only if it satisfies the logic Equation (1). Let $X(i) = f(i)$, $\forall i \in S$.

$$NX(i,k) = X(\delta(i,k)),\ \forall i \in S,\quad k \in X.$$

$$A(i,j) = \begin{cases} 1 & \text{if } X(i) = X(j) \\ 0 & \text{otherwise,} \end{cases} \quad \forall i, j \in S$$

$$B(i,j,k) = \begin{cases} 1 & \text{if } NX(i,k) = NX(j,k) \\ 0 & \text{otherwise,} \end{cases} \quad \forall i, j \in S,\quad k \in X$$

The logic Eq. (1) can be written as

$$(\forall k \in X)(\forall i, j \in S),\ (A(i,j) \supset B(i,j,k)) \tag{2}$$

Since $a \supset b = \neg a \vee b$, logic Eq. (2) can be expressed as

$$(\forall k \in X)(\forall i, j \in S),\ (\neg A(i,j) \vee B(i,j,k)) \tag{3}$$

Equations (1), (2) and (3) are all equivalent and can be written in a high-level computer language such as PL/1 or Algol 60. Figure 1 shows the flow chart for generating all partitions with the substitution property.

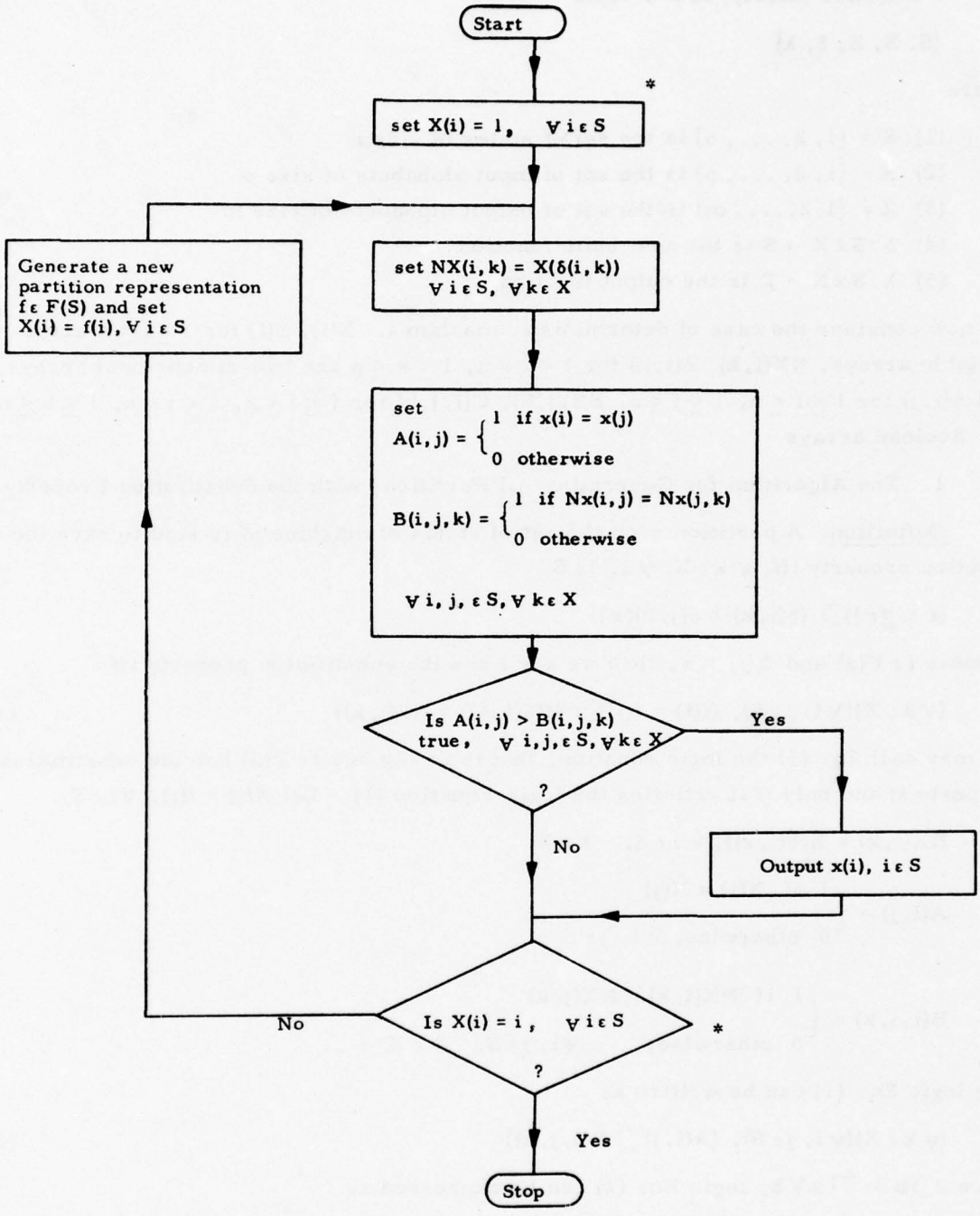


Fig. 1. Flow chart for generating all partitions with substitution property. x(i) = 1, ∀ i ε S, is the universal partition, x(i) = i, ∀ i ε S, is the zero partition.

2. Algorithm for Generating All State-to-State Partition Pairs

Definition:[7] (π, τ) is a State-to-State (SS) partition pair iff, $\forall i \in S$, $\forall k \in X$,

$(i \equiv j(\pi)) \supset (\delta(i,k) \equiv \delta(j,k)(\tau))$

For $f, g \in F(S)$, the partition representations of π and τ, respectively, let $X(i) = f(i)$, $Y(i) = g(i)$ and $NX(i,k) = Y(\delta(i,k))$, $\forall i \in S$, $\forall k \in X$; then f, g represents an SS pair iff

$$(\forall k \in X)(\forall i, j \in S), (X(i) = X(j)) \supset (NX(i,k) = NX(j,k)) \quad (4)$$

The only difference between logic Eqs. (4) and (2) in example 1 is the array $NX(*,*)$ defined by another partition representation g. Figure 2 shows the flow chart for generating all SS partition pairs.

3. Minimal Reduced Sequential Machine

The machine equivalent to machine M having a minimal number of states is called the minimal reduced machine of M.

Definition: Partition π on the set of states is output-consistent iff $\forall k \in X$, $\forall i, j \in S$, $(i \equiv j(\pi)) \supset (\lambda(i,k) \supset \lambda(j,k))$

If partition π is the maximal output-consistent partition with the substitution property, then the quotient machine M/π is the minimal reduced machine.[7] In general, we have to solve the problem for incompletely specified machines; therefore we have to extend the definition of output-consistent partition with substitution property. For the problem of concern in this report, we simply treat the unspecified states or output as don't-care conditions in the sense that the statement $\delta(i,k) \equiv \delta(j,k)(\pi)$ is always true for any π if either $\delta(i,k)$ or $\delta(j,k)$ is unspecified; similarly the statement $\lambda(i,k) \supset \lambda(j,k)$ is always true if either $\lambda(i,k)$ or $\lambda(j,k)$ is unspecified. Then we have the extended definition:

Partition π is output-consistent with the substitution property iff $\forall k \in X$, $\forall i, j \in S$,

$(i \equiv j(\pi)) \supset ((\delta(i,k) \equiv \delta(j,k)(\pi)) \vee (\delta(i,k)$ unspecified$) \vee (\delta(j,k)$ unspecified$))$, and

$(i \equiv j(\pi)) \supset (\lambda,(i,k) \equiv \lambda(j,k)) \vee (\lambda(i,k)$ unspecified$) \vee (\lambda(j,k)$ unspecified$))$.

For $f \in F(S)$, the partition representation of π, let $X(i) \supset f(i)$ for $\forall i \in S$.

$$NX(i,k) = \begin{cases} X(\delta(i,k)) & \text{if } \delta(i,k) \text{ specified} \\ n+1 & \text{if } \delta(i,k) \text{ unspecified (note that } n+1 \notin S), \ \forall i \in S, \ \forall k \in X \end{cases}$$

$$Z(j,j) = \begin{cases} \lambda(j,k) & \text{if } \lambda(j,k) \text{ specified} \\ m+1 & \text{if } \lambda(j,k) \text{ unspecified (note that } m+1 \notin Z), \ \forall j \in S, \ \forall k \in X \end{cases}$$

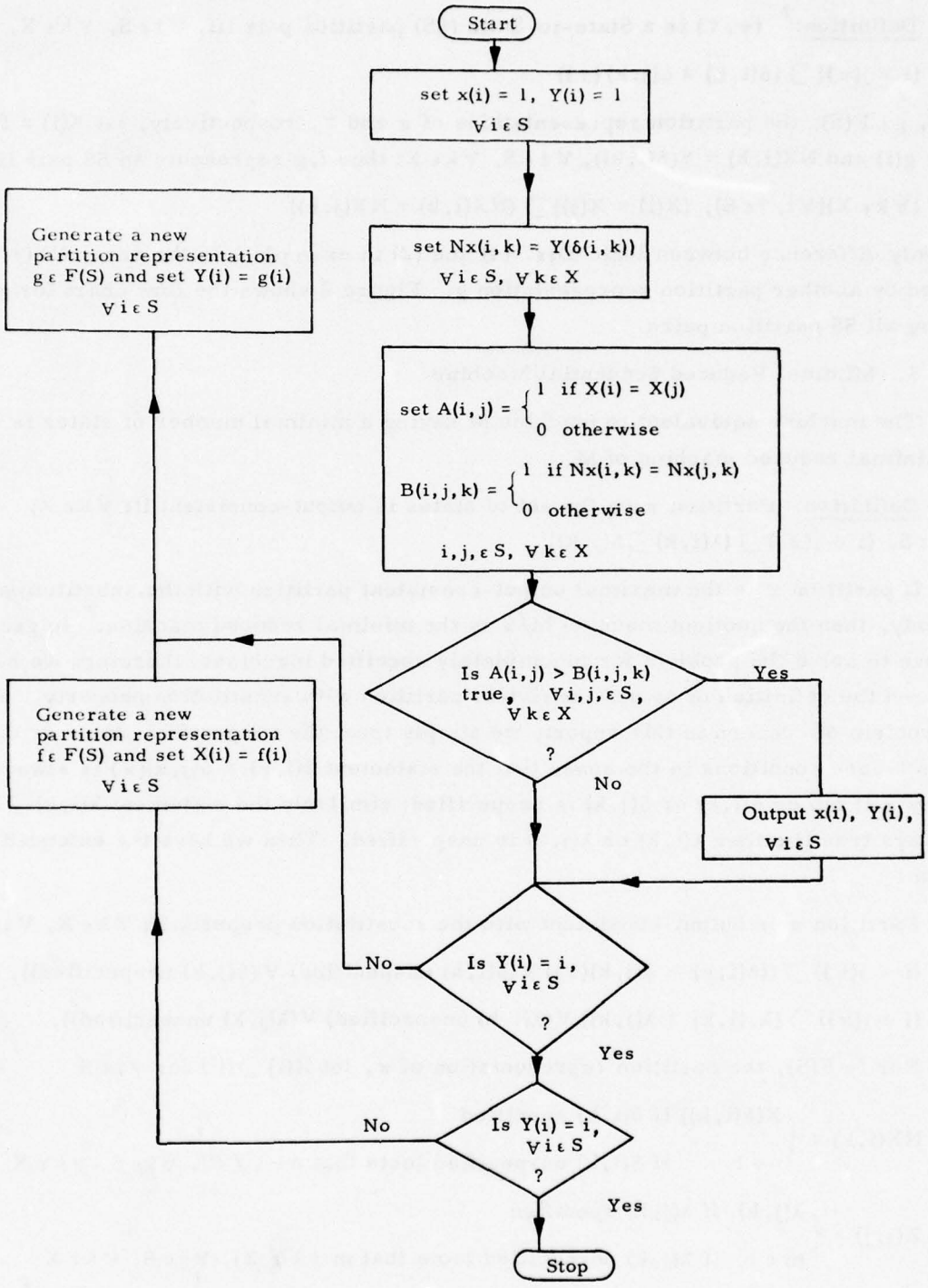


Fig. 2. Generating all SS partition pairs.

$$A(i,j) = \begin{cases} 1 & \text{if } X(i) = X(j) \\ 0 & \text{otherwise, } \forall i, j \in S \end{cases}$$

$$B(i,j,k) = \begin{cases} 1 & \text{if } (NX(i,k) = NX(j,k)) \vee (NX(i,k) = n+1) \vee (NX(j,k) = n+1) \\ 0 & \text{otherwise, } \forall i, j \in S, \forall k \in Z \end{cases}$$

$$C(i,j,k) = \begin{cases} 1 & \text{if } (Z(i,k) = Z(j,k) \vee (Z(i,k) = m+1) \vee (Z(j,k) = m+1) \\ 0 & \text{otherwise, } \forall i, j \in S, \forall k \in Z \end{cases}$$

It follows that f is output-consistent with the substitution property iff

$$(\forall k \in X)(\forall i, j \in S), (A(i,j) \supset (B(i,j,k) \wedge C(i,j,k))) \tag{5}$$

The flow chart is the same as the one in Fig. 1, except that the logic equation is Equation (5). For machine M completely specified, we simply remove the don't-care condition in the definition. Furthermore, since the output-consistent partition is always smaller than or equal to the 1-equivalent partition P_1 (see Ref. 8, p. 287), the initial partition (the universal partition) in the flow chart can be replaced by P_1, which is obtained from the state table of completely specified machine M.

4. Congruences of Semigroup

A finite semigroup $\{G,p\}$ is a set $G = \{1,2,\ldots,n\}$ with an associative binary composition $p: G \times G \rightarrow G$. In the following discussion we will use the terms equivalence relation and partition interchangeably.

Definition: An equivalence relation R(L) over a semigroup G is called a right (left) congruence if $a\,R\,b$ implies $p(a,x)\,R\,p(b,x)$ ($a\,L\,b$ implies $p(x,a)\,L\,p(x,b)$ respectively, $\forall x \in G$. An equivalence relation E over G which is both a right and a left congruence is called a congruence over G; E is a congruence over G, iff $a\,E\,b$, $c\,E\,d$ implies $p(a,c)\,E\,p(b,d)$. The factor structure G/E is called the quotient semigroup of the semigroup G.[5,6]

Let F(G) be the representation for the set of partitions on G. For $f \in F(G)$, f is a right congruence iff, $\forall i,j,k \in G$,

$$(f(i) = f(j)) \supset (f(p(i,k)) = f(p(j,k))) \tag{6}$$

Similarly, f is a left congruence iff, $\forall i, j, k \in G$,

$$(f(i) = f(j)) \supset (f(p(k,i)) = f(p(k,j))) \tag{7}$$

Consequently, f is a congruence iff, $\forall i, j, k, m \in G$,

$$((f(i) = f(j)) \wedge (f(k) = f(m))) \supset (f(p(i,k)) = f(p(j,m))) \tag{8}$$

Let

$$A(i,j) = \begin{cases} 1 & \text{if } f(i) = f(j) \\ 0 & \text{otherwise, } \forall\, i, j \in G \end{cases}$$

then the logic Eqs. (6), (7) and (8) can be simply expressed by

$$(\forall\, i, j, k \in G), A(i,j) \supset A(p(i,k), p(j,k)) \tag{9}$$

$$(\forall\, i, j, k \in G), A(i,j) \supset A(p(k,i), p(k,j)) \tag{10}$$

$$(\forall\, i, j, k, m \in G), (A(i,j) \wedge A(k,m)) \supset A(p(i,k), p(j,m)) \tag{11}$$

Eqs. (9), (10) and (11) respectively.

The flow chart for generating congruences is straightforward.

We have achieved the objective: a unified and systematic approach to construct various algorithms for computing the quotient structures of finite automata as well as other algebraic structures, such that the algorithms can be easily written into any popular high-level computer language. In the following, the procedures for computing the SP partitions for a given machine is illustrated. Other examples are given in Reference 12. Further research might include algorithms for computing the quotient structures of non-deterministic automata[11] and other related disciplines such as transition graphs.

PL/I Program to Compute the SP Partition of a Given Machine

The state table of a given machine M is as follows:

PS	NS				
	x_1	x_2	x_3	x_4	x_5
1	2	1	5	8	3
2	1	2	6	7	4
3	4	3	6	6	1
4	3	4	5	5	2
5	5	6	3	4	7
6	6	5	4	3	8
7	7	8	4	2	5
8	8	7	3	1	6

Machine M

The partitions with substitution property and their lattice are:

$I = \{\overline{1,2,3,4,5,6,7,8}\}$

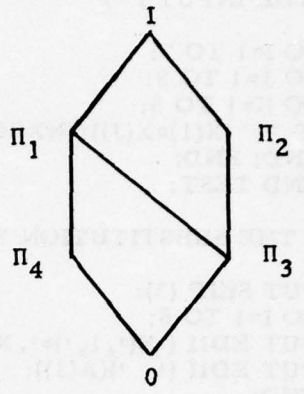
$\Pi_1 = \{\overline{1,2,3,4}, ; \overline{5,6,7,8}\}$

$\Pi_2 = \{\overline{1,2,7,8}, ; \overline{3,4,5,6}\}$

$\Pi_3 = \{\overline{1,2}, ; \overline{3,4} ; \overline{5,6} ; \overline{7,8}\}$

$\Pi_4 = \{\overline{1,4} ; \overline{2,3} ; \overline{5,8} ; \overline{6,7}\}$

$0 = \{\overline{1} ; \overline{2} ; \overline{3} ; \overline{4} ; \overline{5} ; \overline{6} ; \overline{7} ; \overline{8}\}$



Lattice L_M

PL/1 - Program

```
PASP: PROC OPTIONS (MAIN);
      DCL (X(8),NX(5,8)) FIXED DEC (1,0);
      PUT LIST ('THESE ARE PARTITIONS WITH SUBSTITUTION PROPERTY
OF THE GIVEN MACHINE M');
      X(1)=1;
      DO X(2)=1 TO 2;
      DO X(3)=1 TO 3;
      DO X(4)=1 TO 4 WHILE (X(X(3))=X(3));
      DO X(5)=1 TO 5 WHILE (X(X(4))=X(4));
      DO X(6)=1 TO 6 WHILE (X(X(5))=X(5));
      DO X(7)=1 TO 7 WHILE (X(X(6))=X(6));
GET:  DO X(8)=1 TO 8 WHILE (X(X(7))=X(7));
      P=X(8);
      IF X(P)^=X(8) THEN GOTO GIVUP;

/* GENERATE A NEW PARTITION REPRESENTATION X(I) WHICH SATISFIES
THE CONTRACTION AND IDEMPOTENT CRITERIA */

      NX(1,1)=X(2);     NX(1,2)=X(1);
      NX(1,3)=X(4);     NX(1,4)=X(3);
      NX(1,5)=X(5);     NX(1,6)=X(6);
      NX(1,7)=X(7);     NX(1,8)=X(8);
      NX(2,1)=X(1);     NX(2,2)=X(2);
      NX(2,3)=X(3);     NX(2,4)=X(4);
      NX(2,5)=X(6);     NX(2,6)=X(5);
      NX(2,7)=X(8);     NX(2,8)=X(7);
      NX(3,1)=X(5);     NX(3,2)=X(6);
      NX(3,3)=X(6);     NX(3,4)=X(5);
      NX(3,5)=X(3);     NX(3,6)=X(4);
      NX(3,7)=X(4);     NX(3,8)=X(3);
      NX(4,1)=X(8);     NX(4,2)=X(7);
      NX(4,3)=X(6);     NX(4,4)=X(5);
      NX(4,5)=X(4);     NX(4,6)=X(3);
      NX(4,7)=X(2);     NX(4,8)=X(1);
      NX(5,1)=X(3);     NX(5,2)=X(4);
      NX(5,3)=X(1);     NX(5,4)=X(2);
      NX(5,5)=X(7);     NX(5,6)=X(8);
      NX(5,7)=X(5);     NX(5,8)=X(6);
```

```
/* NX(I,J) IS THE NEXT STATE OF STATE J
UNDER THE INPUT I */

TEST: DO I=1 TO 7;
      DO J=1 TO 8;
      DO K=1 TO 5;
      IF ^(^(X(1)=X(J))!(NX(K,1)=NX(K,J))) THEN GO TO GIVUP;
      END; END;
      END TEST;

/* TEST THE SUBSTITUTION PROPERTY OF THE PARTITION */

      PUT SKIP (3);
      DO I=1 TO 8;
      PUT EDIT ('X(',I,')=',X(I))(A(2),F(1),A(2),F(1));
      PUT EDIT ('  ')(A(3));
      END;
GIVUP: END GET;
      END; END; END; END; END; END;
      END PASP;
```

THESE ARE PARTITIONS WITH SUBSTITUTION PROPERTY OF THE GIVEN MACHINE M

X(1)=1	X(2)=1	X(3)=1	X(4)=1	X(5)=1	X(6)=1	X(7)=1	X(8)=1
X(1)=1	X(2)=1	X(3)=1	X(4)=1	X(5)=5	X(6)=5	X(7)=5	X(8)=5
X(1)=1	X(2)=1	X(3)=3	X(4)=3	X(5)=3	X(6)=3	X(7)=1	X(8)=1
X(1)=1	X(2)=1	X(3)=3	X(4)=3	X(5)=5	X(6)=5	X(7)=7	X(8)=7
X(1)=1	X(2)=2	X(3)=2	X(4)=1	X(5)=5	X(6)=6	X(7)=6	X(8)=5
X(1)=1	X(2)=2	X(3)=3	X(4)=4	X(5)=5	X(6)=6	X(7)=7	X(8)=8

National Science Foundation
GK 43437X

T. Lee and E.J. Smith

REFERENCES

1. G. Hutchinson, "Partitioning Algorithms for Finite Sets," Comm. of ACM, 6, No. 10 (October 1963).
2. G. Ehrlich, "Loopless Algorithms for Generating Premutations, Combinations, and other Combinatorial Configurations," J. ACM, 20, No. 3 (July 1973).
3. G. Ehrlich, "Algorithm 477, Generator of Set-Partitions to Exactly R Subsets," Collected Algorithms from CACM.
4. G. Birkhoff and S. MacLane, "A Survey of Modern Algebra," (New York: Macmillan Company, 1948).
5. N. Jacobson, "Basic Algebra I," (San Francisco: W.H. Freeman and Co., 1974).
6. A. Ginzburg, "Algebraic Theory of Automata," (New York: Academic Press, 1968).
7. J. Hartmanis and R.E. Stearns, "Algebraic Structure Theory of Sequential Machines," (Englewood Cliffs, N.J.: Prentice Hall, 1966).
8. Z. Kohavi, "Switching and Finite Automata Theory," (New York: McGraw Hill, 1970).
9. M.B. Wells, "Elements of Combinatorial Computing," (Oxford: Pergamon Press, 1971).
10. C. Berge, "Principles of Combinatorics," (New York: Academic Press, 1971).
11. R.T. Yeh, "On Relational Homomrphisms of Automata," Info. and Control, 13 (1968).
12. T. Lee and E.J. Smith, "On the Partitioning Representation for Finite Sets and Its Application," Research Report POLY EE/EP 76-009 (May 1976).

A SEMI-AUTOMATIC DESIGN VERIFICATION SYSTEM FOR COMMUNICATIONS LINE CONTROL PROCEDURES

E.J. Lancevich and E.J. Smith

The present report illustrates the use of a communications protocol design language (CPDL) as a powerful tool in a semi-automatic design-verification system. The application of CPDL to the description of line-control procedures is described in another section of the present report.[1]

A. Design Automation of Line Control Procedures

In the following sections, some techniques will be described that may be used in a semi-automated design environment where the final goal is the implementation of line control procedures. We choose to use the word implementation rather than hardware, software or firmware realization, since the state-of-the-art is so rapidly changing in the field. There are many considerations that must be taken into account in an overall design automation environment, such as the ultimate output (what is to be obtained), the cost of developing the system, etc. Our focus here is on verifying the design of line control procedures, that is procedures that check for "correctness" of the line controls.

Consider the block diagram shown in Figure 1. In this diagram, the CPDL description (program) is fed to a computer, indicated by the dotted enclosure which

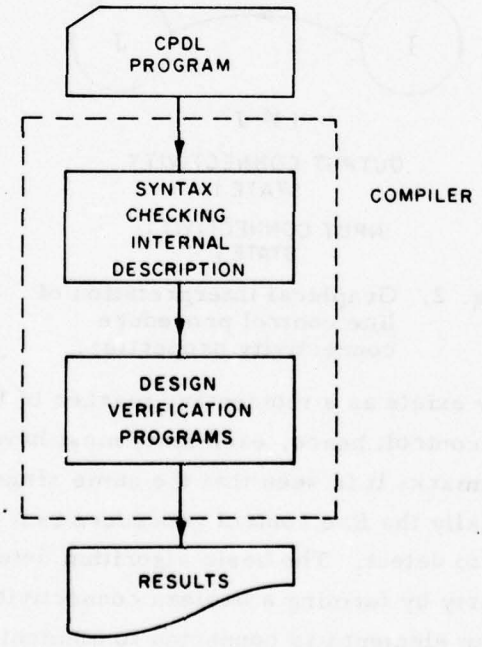


Fig. 1. Elements of a design verification system using CPDL as a source language.

contains a syntax checker and internal description generation block. The syntax checker is simply used to check the structural accuracy of the code while the internal description generation block is used to generate a set of data, describing CPDL in a way which can be used as input to the design verification block. In a broad sense, the syntax checker and internal description block is a compiler for the CPDL language, where the output is not executable software, but data useful in design verification. In the following sections we develop some criteria which all correct CPDL programs must satisfy and we give algorithms for the checking of these criteria.

B. Verification of Line Control Procedures

All line control procedures must possess certain properties in order to be capable of execution within some implementation. Such properties are presented in the following subsections along with the algorithms for their detection.

1. Output Connectivity

The property of output connectivity is that for any state there must exist at least one state transition for which that state is the source and in which it is not the destination. This essentially says that, if viewed in terms of a state diagram, a state must be linked to at least one state other than itself. This situation is illustrated in Figure 2. The condition is consistent with the operational or dynamic environment of the



$I \neq J$

OUTPUT CONNECTIVITY
STATE i

INPUT CONNECTIVITY
STATE j

Fig. 2. Graphical interpretation of line control procedure connectivity properties.

system, since the state only exists as a momentary marker of the time sequence of signals that are driving the control; hence, each state must have another state as its destination. From these remarks it is seen that the same situation applies to all states in the procedure; so essentially the line control procedure is a closed system. This property is extremely easy to detect. The basic algorithm determines whether the system possesses the property by forming a boolean connectivity matrix C, with elements c_{ij}, where $c_{ij} = 1$ when element i is connected to element j and $c_{ij} = 0$ otherwise.

This is most easily achieved by assigning a number to each state in the symbol table of the compiler in a first pass, and in the second pass, using the number assigned to make an entry in the matrix when the state transition field of the CPDL statement is rescanned. For example, if the state transition field of the statement contained the entry INIT,RCV, on the first pass the compiler creates a unique number for the symbol INIT, say 2 and for RCV, say 7 and the element c_{27} is then set equal to one. Having the C matrix, the following semi-formal algorithm checks for the connectivity, where N is the number of distinct states in the line control procedure.

Algorithm

Input: Connectivity matrix C, size N by N

Output: Marked states indicating which states are Output Connected

```
1. do i=1 to N
2. do j=1 to i-1, j=i+2 to N
       if c_ij=1 then begin mark element "i"
                           goto 2
                      end
3. stop
```

The underlined words are keywords in the semi-formal programming language. The outer do loop controls the stepping through the rows of the matrix, while the inner do loop marks element i if it is connected to another element, not itself. The end closes the do statements and the begin statement. Note that connectivity is just a necessary condition for the procedure described by CPDL. It clearly does not guarantee or say anything about whether the line control performs correctly.

2. Input Connectivity

Just as the connection out of a state is necessary, so is input connectivity; that is, for state i, there exists a state j different from i from which a transition must occur. This is apparent from the preceding remarks and additionally from the fact that any state must be attainable from the idle state defined in the program. The condition is tested by forming the transpose of the connectivity matrix C defined in Section 1 and then applying the same test. In this case, the marked output will be the set of unreachable states. These algorithms may be programmed and used as part of the design verification system. For designs involving many states, the ones of most interest, errors such as the ones detected by the two described algorithms are bound to occur, so that even though the tests appear to be rather simple, they are essential. The tests for the connectivity properties are performed on the static description of the line control procedure, are straightforward,and moreover, lead to the conclusion from

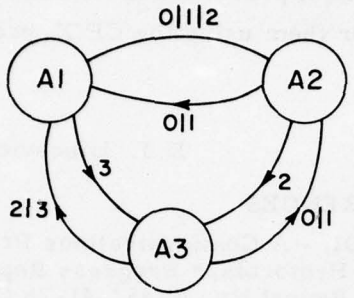
the switching theoretic point of view that the graph is strongly connected; that is, any state may be reached from any other state under some sequence of inputs. While we have no immediate use for this property, it may be of use in future investigations.

3. Completeness

The property of completeness is another property that all line control procedures must possess. Basically, we require that all signals in the same set be accounted for when any signal from that set causes a transition out of a state. For example, if the control is in some state "A" and a transition is caused by the arrival of the character "SOH," where "SOH" is an element of a signal set with four elements, the three other elements of the set must appear on some transition out of state "A." The reason behind this is that we tacitly assume that an error may occur on the transmission line which transforms one element of the set into another and goes undetected in the signal processor. In a deterministic design, there must be a defined action for these errors and thus checking for the occurrence of this phenomenon is a valid design verification function. In the case of feedback signals from the multiplexer, we assume reliable signalling; thus for these signals, no such checking is necessary. In checking for this property, the algorithm is more suitably placed within the context of the block labelled compiler, (the dotted enclosure in Figure 1). The general idea of the algorithm is to evaluate the signal expressions as they are parsed (by the syntax checker) for each current state of the control. This evaluation results in a set of signals. A vector for each signal set defined within the line control procedure, i.e., appearing in the definitions section of the CPDL program, is created and initially set to zero. At each statement associated with the same state in the program, the result of the signal expression evaluation is used to mark the signals that appeared for the particular statement. After passing over the entire input program for the particular state, the vectors are checked. If the signal set vector has all ones or all zeroes, the signal set has been accounted for (all ones) or was not present in the transition (all zeroes). If neither of these conditions is met, then in those vectors that contain the combinations of zeroes and ones, the signals corresponding to the zeroes have not been accounted for. The checking procedure is repeated for all states.

The algorithm is illustrated by an example, instead of a more formal step-by-step procedure.

The state transition graph for the example is shown in Fig. 3, along with the corresponding CPDL program. In the interest of clarity, no processing functions are included. The only signal set defined is the set of integers 0, 1, 2, 3. A four-element vector is set initially to zero. When the program implementing the check algorithm



CPDL PROGRAM

```
DEFINITIONS
SIGNAL (0,1,2,3)
START = A1
CONTROL
(0|1|2)/ A1, A2/.....
      3 / A1, A3/.....
      2 / A2,A3/.....
    0|1)/ A2,A1 /.....
    2|3)/ A3,A1 /.....
    0|1)/ A3,A2/.....
```

Fig. 3. Example for completeness check algorithm.

reads in the first line of the control section, an association is made between the vector and the state appearing on the left-hand side of the state transition field. In the first line (or statement), the signal expression evaluates to the set 0, 1, 2, so those elements of the vector corresponding to those signals are set to one; for example, the vector might contain, after the operation has been performed, 1, 1, 1, 0. The program reads the next input statement. Since the state is A1, the signal expression is evaluated and the new signal, 3, is noted by entering a one in the last vector element. In each of the steps, the input lines of the program are marked indicating that they have been operated upon. When all lines have been marked, the check algorithm terminates.

Since the next four statements do not contain the state A1 in the left hand side of a state transition, they are skipped over by the check program. At the conclusion of the first pass over the input program, all signals are accounted for at A1.

A second iteration of the procedure is then started. During the second pass, the signal set vector is set to zero at the outset, and the check program starts to read the CPDL program again. Since the first two statements are marked, the signal set vector is now associated with state A2. In a manner similar to that described above, the lines associated with A2 are processed, marked, and at the conclusion of the pass, the vector contains 1, 1, 1, 0 indicating that at A2, a transition has not been provided for the signal 3. The algorithm terminates after processing state A3.

It should further be noted that this algorithm may be used to determine whether the program contains two or more transitions from the state caused by the same input signal, by checking at the step where the vector is set. If the bit associated with the signal is already a one, the signal has already defined a transition.

C. Discussion

The work reported demonstrates the feasibility of a useful design automation technique which uses CPDL as an input language. The constraints imposed by the communications environment on the line control procedure have led to the inference of a basic set of properties which each line control procedure must satisfy and to the effective procedures described to check for them using the CPDL program as input to the algorithms.

National Science Foundation
ENG 73-04232

E.J. Lancevich and E.J. Smith

REFERENCES

1. E.J. Lancevich and E.J. Smith, "CPDL - A Communications Protocol Design Language: Application to Full-Duplex Protocols," Progress Report No. 41 to JSTAC, Polytech. Inst. of New York, Report No. R-452.41-76 (1976). Also published as Research Report POLY EE/EP 76-010 (June 1, 1976).

2. E.J. Lancevich, "The Systematic Design of Communication Line Control Procedures," Ph.D. Dissertation, Polytech. Inst. of New York (June 1976).

CPDL - A COMMUNICATIONS PROTOCOL DESIGN LANGUAGE: APPLICATION TO FULL-DUPLEX PROTOCOLS

E.J. Lancevich and E.J. Smith

A communications protocol design language (CPDL) was previously developed for the precise definition of simplex-type communication protocols.[1] Additional work[2] summarized in the present report has shown that the language can be effectively extended to the more complex case of full-duplex protocols as well. The extended BNF description of CPDL is given. An included example illustrates both the CPDL description of a full-duplex protocol and its corresponding state-diagram representation.

A. Full-Duplex Protocols

The primary distinguishing factor between simplex, half-duplex, and full-duplex line controls is the parallelism that exists in the latter case, since transmission takes place in both directions on the line at the same time. In most cases of interest, however, full-duplex protocols rely on sequence numbering in each direction to maintain synchronism between sending and receiving processes at the switch. Sequence numbers are usually generated by a modulo k numbering scheme, in which k is the largest number of messages (assuming equal message lengths) that may be propagating in both directions on the communications line. As an example, consider a scheme in which the blocks 1000 bits long are transmitted over a channel with capacity resulting in a one way delay of 1 second. Then the equivalent length of the transmission media (round trip) is 4800 bits. Thus five messages (total) may reside in the channel at one time, so that the acknowledgement signal for the first message sent by one end arrives at the other end at the same time the fifth message is being transmitted, in the worst case; so that for this example k = 5. Gray[3] presents an excellent survey of full-duplex protocols and gives strong arguments favoring the use of independent sequence numbering in each direction.

We apply our proposed model to just such a case. For simplicity, assume that messages have the format shown in Figure 1. The character SOM refers to the start of the message. This is followed by a fixed-length indicator transmit sequence number, a fixed-length information block and a fixed length receive number. Again to simplify, assume there are no checking characters. Assume also that the length of each sequence number is 3 and the length of the information field is 50.

Both ends transmit and receive simultaneously with two independent sequence numbers; when no errors occur, the transmit sequence number received at one end of the connection will be placed into the received sequence number at the same connection and then transmitted back to the original send (as a positive acknowledgement). However, if the original message is garbled, the receiver transmits the last properly

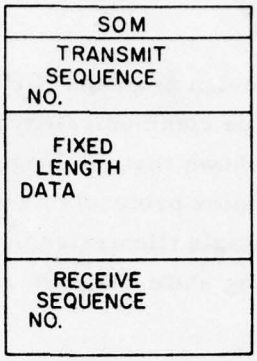


Fig. 1. Message format for the full-duplex example.

received sequence number and destroys the message. This effectively is a negative acknowledgement signal and causes the original message plus the next set of messages (until the time the negative acknowledgement of original message is received)to be misinterpreted.

To illustrate, suppose k= 5, and A sends to B. Let the messages be labelled M_0, M_1, M_2, M_3, M_4. If M_2 is received in error by B, then in two consecutive transmissions, A receives ones in the corresponding number field. If A has already transmitted M_3 and M_4, they will be received in error at B (out of sequence) and discarded from the system. Thus in this scheme, M_2, M_3 and M_4 must be retransmitted in order to maintain channel synchronism.

The finite-state graph of the receiver describing this scheme is shown in Fig. 2 for one end of the connection and the sender transition graph is shown in Figure 3. The receiver responds to a start-of-message signal initially and proceeds to collect the first three digits. If the transmitted sequence number is equal to the expected sequence number then the expected number is increased by one mod 5 and the signal vtc is sent via the cause primitive of CPDL (see Figure 5).

If the number transmitted is not equal to the proper number, the signal invtc causes a transition to the error state (ERROR) and the rest of the transmission is ignored. In this diagram, we assume that the signals invtc and vtc occur before the next significant transmission.

From the error state, once the last transmission is received, operations are executed within the multiplexer which result in the next outbound transmission to have placed in its receive field the sequence number of the last valid block received. The transmission of the block is executed the sender state diagram and the ready signal generated by the cause primitive at the receiver.

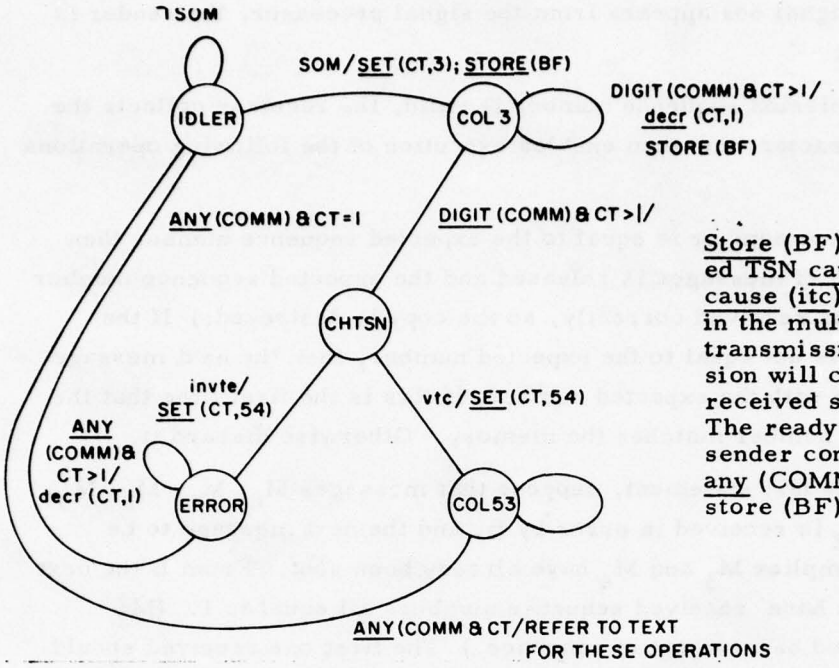


Fig. 2. Receiver state graph for the full-duplex example.

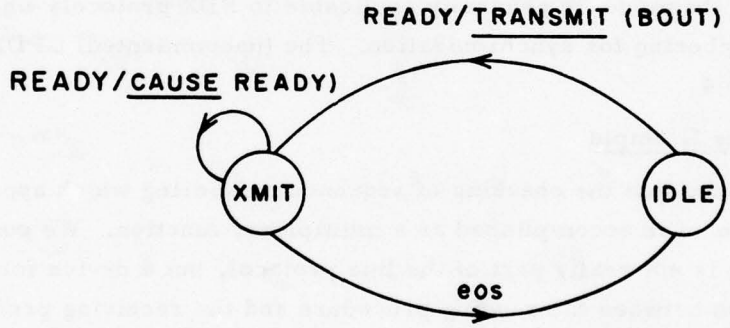


Fig. 3. Sender state graph for the full-duplex example.

If the transmission sequence number is invalid, the multiplexer processing proceeds to load the last valid received sequence number into the next outbound message and prompt the sending procedure for the line with a ready signal. If the sending process is idle, it transmits the next outbound message. Otherwise, it prompts the multiplexer to reenable it to send again if it is already transmitting another message (in effect a waiting loop for the transmission). This is depicted in Fig. 3, the sender

state diagram. When the signal eos appears from the signal processor, the sender is again enabled to retransmit.

Finally, if the transmission sequence number is valid, the receiver collects the rest of the transmitted characters and then enables execution of the following operations in the multiplexer.

If the received sequence number is equal to the expected sequence number then that message slot in the list of messages is released and the expected sequence number incremented by one. (It was received correctly, so the copy is destroyed.) If the received sequence number is not equal to the expected number, then the next message to be transmitted is the one with the expected number, if this is the first time that the same "last valid received" number matches the memory. Otherwise discard it.

To further clarify this last statement, suppose that messages M_0, M_1, M_2, M_3, M_4 are sent from A and M_2 is received in error by B, and the next message to be transmitted is M_0, which implies M_3 and M_4 have already been sent. From B the next three return messages will have received sequence numbers all equal to 1. (M_2 garbled, M_3 and M_4 rejected because out of sequence.) The first one received should cause M_2 to be transmitted next and the next two should be ignored (up to a maximum of 5 and then the process should be restarted), since these are the result of M_3 and M_4 being ignored at the receiver due to the garbling of M_2. This completes the example which indicates that the model is generally applicable to FDX protocols which use double sequence numbering for synchronization. The (uncommented) CPDL descriptions are shown in Figure 4.

B. Comments on the Example

It should be noted that the checking of sequence numbering which apparently is part of the line protocol is accomplished as a multiplexer function. We contend that the sequence numbering is not really part of the line protocol, but a device for maintaining local synchronization between the sending procedure and the receiving procedure. The sequence numbering merely serves as a device to link the two procedures.

C. CPDL-BNF Description

The latest version of CPDL is defined by the BNF description shown in Figure 5. As more experience is gained in its use, it will be modified and enhanced.

SENDER PROCEDURE

```
definitions
signal     COMM (SOM, A, B, ···, Z, 0, 1, ···, 9)
counter    CT
buffer     BF
start = IDLER
control
 SOM/IDLER, IDLER/nop;
SOM/IDLER, COL3/set (CT, 3);store (BF);
digit(COMM)&CT> 1/COL3, COL3/store(BF);decr(CT, 1);
digit(COMM)&CT=1/COL3, CHTSN/store(BF);     For other operations
                                            refer to Fig. 2 and
                                            Section A of text.
vtc/CHTSN, COL53/set(CT, 54);
any(COMM)&CT> 1/COL53, COL53/store(BF);decr(CT, 1);
any(COMM)&CT=1/COL53, IDLER/store(BF);      Refer to Section A.
invtc/CHTSN, ERROR/set(CT, 54);
any(COMM)&CT  1/ERROR, ERROR/nop;
any(COMM)&CT=1/ERROR, ERROR/nop;
```

RECEIVER PROCEDURE

```
definitions
signal     COM (ready, eos)
buffer     BOUT
start = IDLE
ready/IDLE, XMIT/transmit(BOUT);
ready/XMIT, XMIT/cause(ready);
eos/XMIT, IDLE/nop;
```

Fig. 4. CPDL programs for sender and receiver for full-duplex example.

<PROGRAM> → definitions <DLIST> control <SLIST>
<SLIST> → <SLIST><STMNT> |< STMNT>
<STMNT> → <EEXPR> /<FST> , <TST> /<EXELT>
<FST> → <IDENTIFIER>
<TST> → <IDENTIFIER><IDLST>
<IDLST> → (<IDENTIFIER> {, <IDENTIFIER> })
<EXELT> → <FUNC> ; <EXELT> |<FUNC> ;
<DLIST> → <DLIST><DEFINITION> |<DEFINITION>
<DEFINITION> → counter <IDLST> |buffer <IDLST>
<EEXPR> → <SETEXPR><CNTEXPR>
<CNTEXPR> → <CNT><RLOP><INTEGER> {& <CNTEXPR>}
<CNT> → <IDENTIFIER>
<INTEGER> → <INTEGER><DIGIT> |<DIGIT>
<RLOP> → < |≤ |= |> |≥ |≠
<SETNAME> → <IDENTIFIER> | ascii | ebcdic | baudot | transcode | 4of8
<IDENTIFIER> → <LETTER><ALPHAMERIC><IDENTIFIER>
<ALPHAMERIC> → <LETTER> |<DIGIT>
<LETTER> → a|b|c|d|e|.....y|z|A|B|C ..... Y|Z
<DIGIT> → 0|1|2|3|4|5|6|7|8|9
<SETEXPRESSION> → <SETTRM> '1' <SETEXPRESSION> |<SETTRM>
<SETTRM> → <SETFCT> & <SETTRM> |<SETFCT>
<SETFCT> → '(' SETEXPRESSION ')' |<ELEMENTEXP> |
<UNARYOP> <SETEXPRESSION>
<UNARYOP> → any | not | alpha | digit
<ELEMENTEXP> → <ELN> ‖ <ELEMENTEXP> |<ELN>
<ELN> → <IDENTIFIER>
<DEFINITION> → signal <IDLIST>

Fig. 5. BNF definition of CPDL.

National Science Foundation
ENG 73-04232

E.J. Lancevich and E.J. Smith

REFERENCES

1. E.J. Lancevich and E.J. Smith, "CPDL -- A Communications Protocol Design Language," Progress Report No. 40 to JSTAC, Polytech. Inst. of New York, Report No. R-452.40-75, 448-453 (September 1975).

2. E.J. Lancevich, "The Systematic Design of Communications Line Control Procedures," Ph.D. Dissertation, Polytech. Inst. of New York (June 1976). (Also published as Research Report POLY EE/EP 76-010 (June 1, 1976).)

3. J.P. Gray, "Line Control Procedures," Proc. IEEE, 60, No. 11 (November 1972).

ON A TECHNIQUE FOR DYNAMIC ROUTING

R.R. Boorstyn and A. Livne

Many computer communication networks use dynamic routing schemes to compensate for input traffic variations, to respond to changes in topology, and to take advantage of temporary changes in loading in different paths. These adaptive schemes are complex and are usually chosen and verified by extensive simulations. Invariably during the design of a network they are replaced by analytically tractable non-dynamic (static) schemes. We present here a dynamic routing scheme for which we have been able to derive approximate analytical models.[1,2] Furthermore, we can establish the efficiency of this scheme, especially in heavily loaded situations.

A typical static routing scheme would operate as follows. Consider as separate commodities the messages originating at a particular node and destined for a second node in the network. There are usually several good paths connecting these nodes. The static routing scheme would specify the optimum proportion of traffic to be routed over each path. Efficient algorithms exist for design of this type of routing.

We can identify one particular problem with this approach. Although good paths are indeed found, any node essentially operates as a collection of single server queues -- one queue for each outgoing branch. Considering the node as a queue with several potential servers this is not an efficient manner of operation. Indeed if a node had k outgoing branches and was operated as a queue with k servers then when heavily loaded the time delay would be reduced by a factor of k. Conversely the throughput can be increased.

If the node was operated as suggested the above messages would wander aimlessly through the network and the total performance would be abysmal. Our approach is to retain the good paths for commodities and yet still get the benefit of the faster performance at the node.

Briefly our scheme is as follows. Consider a node as a single queue with several servers (output channels). For a particular commodity, i.e., a message with a certain destination, the use of some of these servers would cause the messages to be sent along "bad" paths -- either too long or too heavily loaded. Thus for each commodity we specify a subset of the output channels as allowable and permit the message to use any allowable channel according to some discipline. Each commodity appearing at the node has its own allowable set of channels. These restrictions force messages to "good" paths and constitute our routing strategy.

We have found that the node retains its performance advantage as a multiple server queue as long as a modest amount of the traffic has choice. We have developed

approximate and fairly accurate analytical models to calculate the performance of these nodes.

We have also developed analytical methods to imbed these nodes in a network and to calculate the overall performance of the network. Basically we find that if the average number of output channels per node is k, then the time delay for messages in a heavily loaded network can be reduced by as much as a factor of k when this dynamic routing is used.

National Science Foundation
ENG73-04232

R.R. Boorstyn and A. Livne

Joint Services Technical Advisory Committee
F44620-74-C-0056

REFERENCES

1. A. Livne and R.R. Boorstyn, "On a Technique for Adaptive Routing," to be presented at IEEE National Telecommunications Conference, Dallas, Texas (November-December 1976).
2. A. Livne, "Dynamic Routing in Computer Communication Networks," Ph.D. Dissertation, Polytechnic Institute of New York, Brooklyn, N.Y. (June 1976).

CHANNEL ASSIGNMENTS FOR CELLULAR MOBILE TELECOMMUNICATIONS SYSTEMS

R.R. Boorstyn and R.J. Pennotti

A cellular mobile telecommunications system[1] uses base stations distributed throughout a metropolitan area to link mobile customers to the land telephone network. Each base is equipped with enough radios to serve customers within the area (cell) for which it is responsible. A radio channel must be assigned to each radio in the system; the same channel may be assigned to more than one radio as long as the radios using any one channel are dispersed sufficiently that a mobile using the channel in one cell is not unduly interfered with by all other mobiles using the same channel. This constraint can be formalized as a minimum allowable co-channel reuse distance in units equal to the cell radius, since the required co-channel separation in a system is directly proportional to the radius of the cells.

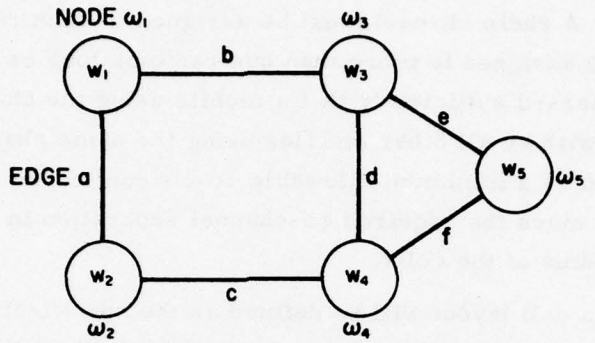
For our purpose, a cell layout will be defined as the specification of cell boundaries throughout a system and the assignment of weights (equal to the number of radios required) to the resulting cells.

The Channel Assignment Problem (CAP) is "Given a cell layout and required co-channel separation distance, how can channels be assigned to cells so as to minimize the total number of channels used?"

A. The Model

Radio interference constraints can be represented by a graph, $G(V, E)$, with a vertex for each radio and an edge between any two vertices corresponding to radios which cannot use the same channel. We will call this graph the system micrograph for reasons which will become clear shortly. The channel assignment problem reduces on the system micrograph to a well-known graph-theoretic problem: that of finding the chromatic number of the graph. This problem has been studied extensively, and a large number of bounds and coloring algorithms have been developed in terms of the connective properties of the graph. Unfortunately, the computational effort necessary to apply these results grows very quickly with the number of vertices in the graph to which they are applied. A graph with 100 vertices is extremely large for consideration of either the chromatic number itself or reasonably tight bounds on the chromatic number. We, on the other hand, are interested in cellular systems which have thousands of radios, and so find ourselves in the familiar position of having at our disposal theoretical results which are not directly applicable to our problem. In order to overcome these computational difficulties, we have developed another graph-theoretic model which takes advantage of the cellular structure of the systems in which we are interested. This model we call the system macrograph, $M(\Omega, \gamma, W)$; it is a graph $G'(\Omega, \gamma)$ with weighted

vertices. (See Figure 1.) W is the set of vertex weights. The vertex set, $\Omega(M)$, is the set of cells in the system. Given ω_i, $\omega_j \in \Omega$, we include (ω_i, ω_j) as an edge (i.e., a member of γ) if and only if cell i interferes with cell j. The weight w_i assigned to vertex ω_i is the number of channels required in cell i.



$\Omega(M) = (\omega_1, \omega_2, \omega_3, \omega_4, \omega_5)$
$\gamma(M) = (a, b, c, d, e, f)$
$W = (w_1, w_2, w_3, w_4, w_5)$

Fig. 1. A macrograph, $M(\Omega, \gamma, W)$.

The system macrograph is a much more compact representation of a cellular system than is the micrograph. It replaces all the vertices representing radios in the same cell with one weighted macronode, taking advantage of the fact that the interference constraints on all radios in the same cell are identical.

The correspondence between G and M is evident. G contains a w_i-clique for each vertex $\omega_i \in \Omega(M)$ (see Figure 2). An edge $(\omega_i, \omega_j) \in \gamma(M)$ represents $w_i \times w_j$ edges in G, fully connecting the w_i-clique from macronode ω_i to the w_j-clique from macronode ω_j. The weight of a node in M is thus the number of nodes in G which it represents.

On the macrograph, the channel assignment problem becomes the problem of assigning w_i distinct positive integers (or colors or channels) to each vertex (cell) of M so that the same integer is not assigned to any two adjacent vertices of M and the total number of assigned integers is minimum. We can use the relationship between macrographs and micrographs to apply known results about the chromatic numbers of graphs to the channel assignment problem in a computationally tractable way. We will illustrate this technique with two well-known bounds on chromatic number in the next section. In subsequent sections we will present a method of partitioning a macrograph into simpler macrographs to compute an upper bound on the optimum CAP solution, and develop a reduction theorem which allows us to eliminate from consideration many of the cells in a typical channel assignment problem.

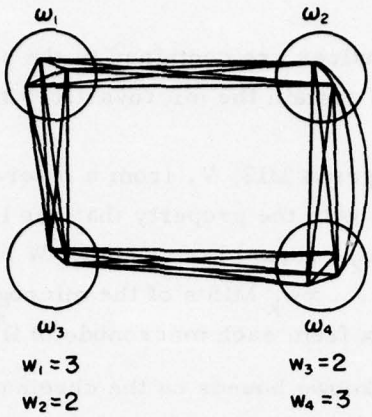


Fig. 2. A macrograph and its corresponding micrograph.

B. The Method

Before developing bounds on the minimum number of channels necessary for the CAP, we must further specify the relationship between macrographs and micrographs.

The first theorem along these lines relates the cliques of the micrograph to the cliques of the macrograph. We will see that there is a simple one-to-one correspondence between the two.

Theorem 1: A set of macronodes is a clique of the macrograph if and only if the micronodes contained* in them form a clique of the micrograph. The correspondence between micrograph and macrograph cliques is therefore one-to-one.

The significance of Theorem 1 becomes evident immediately. If we need to enumerate the cliques of the micrograph we need only enumerate those of the macrograph. The size of a micrograph clique is just the sum of the weights of the nodes in the corresponding macrograph clique. When we speak of the largest micrograph clique, we are speaking of the densest (highest weight) macrograph clique.

A similar theorem can be posed relating macrograph and micrograph maximal independent sets, the major difference being that the correspondence is no longer one-to-one. Any set consisting of a microvertex from each macrovertex of a macrograph MIS is a micrograph MIS. Thus a macrograph MIS corresponds to many equivalent micrograph MIS's.

---

*We will often speak of the micronodes contained in a macronode. By this we will mean those micronodes representing radios in the cell represented by the macronode.

Theorem 2: A set of microvertices is a maximal independent set (MIS) if and only if

(1) no two of the microvertices are contained in the same macronode, and

(2) the macronodes which contain the microvertices form a MIS of the macrograph.

Note that to form a micrograph MIS, V, from a macrograph MIS, Ω^*, we can choose for V any set of vertices with the property that one is selected from each node of Ω^*. If the nodes in $\Omega^* = \{\omega_1, \omega_2, \ldots, \omega_k\}$ have weights $W = \{w_1, w_2, \ldots, w_k\}$ then the MIS Ω^* corresponds to $w_1 \times w_2 \times \ldots \times w_k$ MIS's of the micrograph, representing all combinations of one microvertex from each macronode of Ω^*.

We now discuss two well-known bounds on the chromatic number of a graph and their relationship to the CAP. The thrust of our effort is to bound the chromatic number considering only macrographic properties.

1. Upper Bound (The Maximum Degree Bound)

The chromatic number, $\chi(G)$, of a graph G, is bounded above by the maximum vertex-degree of the graph plus one. However, the degree of any microvertex contained in a particular macronode is just one less than the sum of the weights of the containing macronode and its neighbors (one less because the vertex is not adjacent to itself). Therefore,

$$\chi(G) \leq \max_{\omega_i \in \Omega(M)} \left\{ \sum_{\omega_j \in \Gamma(\omega_i)} w_j + w_i \right\} .$$

Given the macronode weights and adjacencies, we can compute this bound by computing one sum for each cell of the system.

2. Lower Bound (The Largest Clique Bound)

Another well-known graph-theoretic bound on chromatic number is the lower bound given by the size of the largest clique of the graph. As mentioned in the discussion of Theorem 1, the largest clique of the micrograph corresponds to the densest (highest-weight) clique of the macrograph. Thus, the number of channels necessary to solve the CAP is at least as large as the weight of the densest macrograph clique. Mathematically, we let C be the set of all cliques of a macrograph $M(\Omega, \gamma, W)$ with corresponding micrograph $G(V, E)$. Then

$$\gamma(G) \geq \max_{C_i \in C} \left\{ \sum_{\omega_j \in C_i} w_j \right\} .$$

Note that the bound is stated in terms of the connective properties (cliques) of the macrograph only, and the weights of the macronodes.

It is an important lower bound on the number of channels needed to solve the channel assignment problem satisfactorily; the bound is often met in practical systems. Unfortunately, it is difficult to determine whether or not it will be met in a particular system.

While we have presented here only two very simple bounds, the method we have outlined has been successfully applied to much more complex bounds and to quite a number of heuristic graph-coloring algorithms.[2]

C. Vertical Partitioning

It is evident from the relationships which we have established between macrographs and micrographs that the chromatic number of the micrograph is somehow dependent on the structure of the macrograph. For instance, consider a macrograph with equal weights, w, on all the nodes. For this situation, we have $\chi(G) \leq w \cdot \chi(M)$ where M is the macrograph associated with G. A coloring is obtained for G by defining a chromatic coloring for M and letting each of these macrograph colors represent w distinct colors for the micrograph. This concept can be extended to graphs with nonuniform macronode weights to derive a bound on the chromatic number of any micrograph.

Order the cell weights, $\{w_i\}$, of a system so that $w_1 < w_2 < \ldots < w_\ell$ where the system has k cells and $k \geq \ell$. Further, let β_i be the set of cells with weight w_i. Suppose that we know $\chi(M)$ and assign the first $\chi(M)$ channels in such a way that every cell in the system is assigned one channel, that is, find a chromatic coloring for M and assign one channel to each color class. Assuming that each cell needs more than one channel, we assign the next $\chi(M)$ channels in the same way. We continue until eventually the cell needing the fewest channels has been satisfied. Since this cell had weight w_1, we have now assigned $\chi(M) \cdot w_1$ channels. The cells in set β_1 are now satisfied; consider next a new macrograph with these cells eliminated. Since the remaining cells have been assigned w_1 channels, subtract w_1 from their original weights. If we now find the chromatic number of the new macrograph, we can proceed as we did with the original problem until we have completely satisfied all cells needing w_2 channels. This process can be continued iteratively until the entire system is covered. The assignment which results is obviously an upper bound on the minimum assignment. This is a vertical partition of the original macrograph; we partition the weight of each macronode into parts, one satisfied at each iteration of the above process.

The principle of vertical partitioning is even more general than the above examples suggest. Suppose we have a macrograph $M(\Omega, \gamma, W)$ corresponding to the micrograph $G(V, E)$. Consider k copies of the macrograph differing only in their node weights.

Call them $M_1(\Omega, \gamma, W^{(1)})$, $M_2(\Omega, \gamma, W^{(2)})$, ..., $M_k(\Omega, \gamma, W^{(k)})$ and let them correspond to micrographs $G_1, G_2, \ldots, G_k$ respectively. Furthermore let the $w_j^{(i)} \varepsilon W^{(i)}$ satisfy

$$\sum_{i=1}^{k} w_j^{(i)} = w_j \quad ,$$

$j = 1, 2, \ldots, n$ where n is the number of macronodes in the original macrograph (see Figure 3). We can constrain ourselves to serve each of the k copies with disjoint sets of channels and solve the problems independently. This does not guarantee an optimum solution to the original problem but it provides a bound on the optimum.

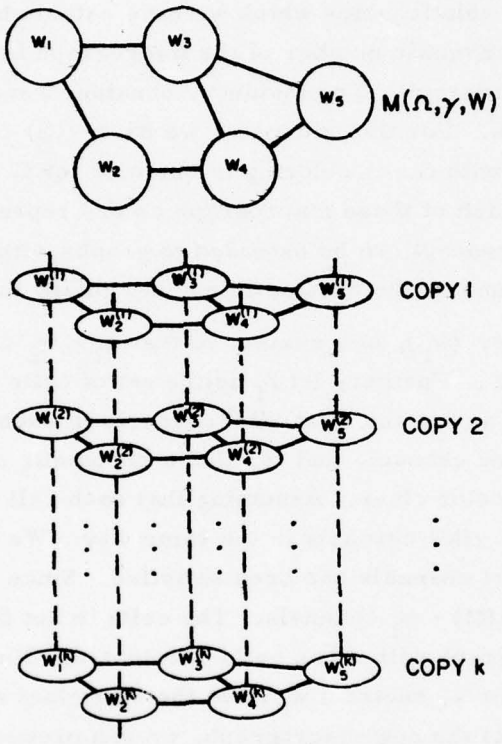


Fig. 3. A vertical partition.

D. The Reduction Theorem

One of the most useful results of this report can be derived from a bound on chromatic number due to Matula.[3] It states that for any cut, $\chi(G)$ is bounded above by the maximum of the chromatic numbers of the resulting subgraphs or the number of edges cut plus one, whichever is larger. Note that if one of the subgraph chromatic numbers is the largest of these quantities, the bound must be met with equality, since the chromatic number of the graph cannot be smaller than the chromatic number of one of its subgraphs.

Now consider a star-cut* of the micrograph. The chromatic number of one of the resulting subgraphs (that with only one vertex in it) can be neglected since it is only equal to one. Moreover, the number of edges cut is equal to the degree of the vertex which is isolated. The bound thus leads to the following assertion (the Reduction Theorem):

> If the chromatic number of the subgraph remaining after a vertex is removed from the graph is greater than the degree of the removed vertex, then the chromatic number of the original graph is equal to the chromatic number of the subgraph.

Suppose we have a lower bound on the chromatic number of the subgraph, e.g., the largest clique. If the degree of the removed vertex is less than this bound, then the vertex need not be considered when searching for the chromatic number of the graph. This theorem can be used iteratively; since each successful step removes a vertex from the graph, the degrees of some of the remaining vertices are decreased. Thus a vertex which does not meet the condition on the first iteration may still be eliminated in a subsequent iteration. Once a subgraph has been reached which cannot be reduced further with this theorem, and it is colored with some number, n, of colors, the eliminated vertices may be colored in the reverse order of their elimination without using any new colors. This is so because the degree of each vertex in the subgraph from which it was removed had to be less than n for it to be eliminated.

So far we have spoken of the Reduction Theorem as it applies to micrographs; but we have stressed all along that the micrograph is an unacceptable model for the CAP because of its size. We must therefore adapt the Reduction Theorem to macrographs in order to take full advantage of it. This can be done through the use of two observations made earlier. The first is that the degrees of all microvertices in a particular macronode are the same. Thus, if one microvertex can be eliminated, then all other microvertices in the same macronode can be eliminated. The second observation is that the degree of a microvertex can be determined from macrographic properties alone; it is one less than the sum of the weights of the containing macronode and its neighboring macronodes. The Reduction Theorem can hence be applied directly to macrographs; through computation of sums of macrograph weights, entire macronodes can be eliminated from consideration in the CAP.

The use of the Reduction Theorem is illustrated through the example in Figure 4. The densest clique consists of macronodes 1 to 3 and gives a lower bound of 25. The degree of each microvertex in macronode 4 is 28 while the degree of each microvertex

---

*A star-cut of a graph separates one vertex from the remainder of the graph.

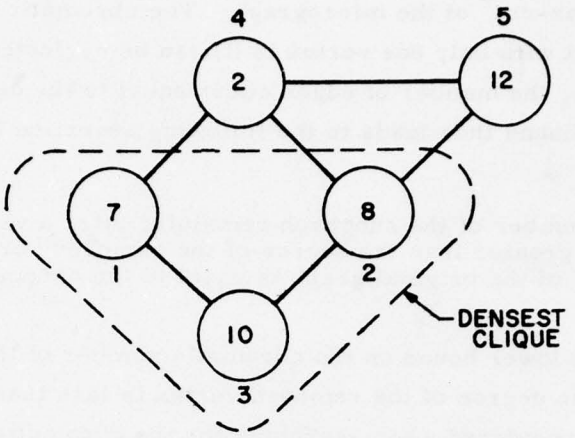


Fig. 4. Example for reduction theorem.

in macronode 5 is 21. Therefore macronode 5 can be eliminated immediately. But the degree of each microvertex in macronode 4 with respect to the remaining micrograph is 16 and so macronode 4 can also be eliminated. Now if the first seven channels are assigned to cell 1, the next eight to cell 2, and the next ten to cell 3, the micrograph remaining after using the Reduction Theorem will have been chromatically colored. We may assign channels 24 and 25 to cell 4 and channels 1 through 7 and 16 through 18 to cell 5 and the solution is complete.

Figures 5 to 7 are cell layouts for a real system at various stages of growth. The

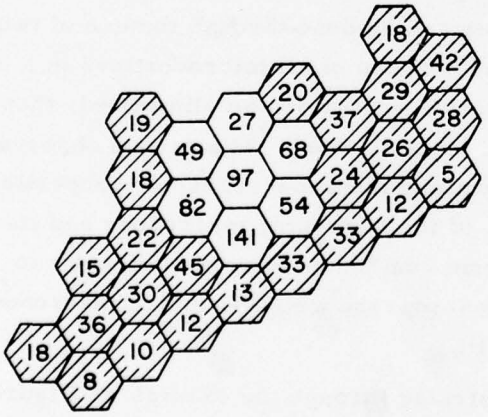


Fig. 5. Cell layout No. 1.

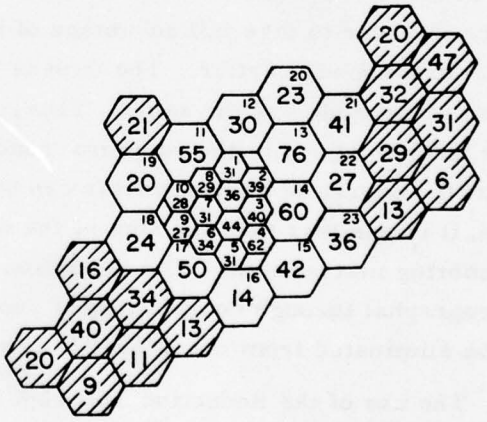


Fig. 6. Cell layout No. 2

first shows the system shortly before a second cell-size is introduced, the second shows the system just after the densest cells have been split and the last shows a mature system close to saturation, given the smallest cell-size and available spectrum. They

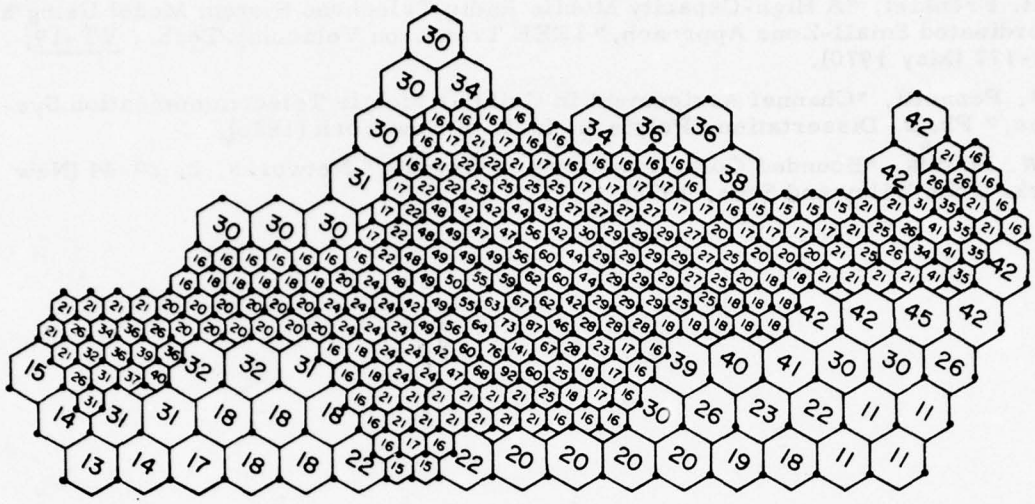


Fig. 7. Cell layout No. 3.

were reduced to 7, 25, and 15 cells respectively. The latter two examples give an indication of the value of the Reduction Theorem in systems with two cell-sizes. The fact that it performs well in these layouts is significant since their lack of symmetry makes them more difficult to analyze than single cell-size layouts. The effectiveness of the Reduction Theorem in the mature system of Fig. 7 is especially encouraging.

E. Summary

We have presented in this report a number of techniques which can be used to make the Channel Assignment Problem in cellular mobile telecommunications systems a tractable problem.

The first technique is simply the development of the macrographic model. But many systems have so many cells that this is not sufficient. The Reduction Theorem allows us to reduce almost any system to its most densely populated core of cells. Once this has been done, the macrographic representations of well-known bounds and algorithms[2] can be used to solve the CAP on these dense cells. As pointed out in the last section, the cells which were eliminated can then easily be assigned channels out of those used for the central core.

Joint Services Technical Advisory Committee
F44620-74-C-0056

National Science Foundation
ENG 73-04232

R.R. Boorstyn and R.J. Pennotti

REFERENCES

1. R.H. Frenkiel, "A High-Capacity Mobile Radio Telephone System Model Using a Coordinated Small-Zone Approach," IEEE Trans. on Vehicular Tech., VT-19, 173-177 (May 1970).

2. R.J. Pennotti, "Channel Assignment in Cellular Mobile Telecommunication Systems," Ph.D. Dissertation, Polytech. Inst. of New York (1976).

3. D.W. Matula, "Bounded Color Functions on Graphs," Networks, 2, 29-44 (New York: John Wiley and Sons, 1972).

QUEUEING ANALYSIS OF INTERNODAL ACKNOWLEDGMENT IN COMPUTER-COMMUNICATION NETWORKS

R. Boorstyn and J. Stark

In order to provide reliable communications in a computer network, messages must be verified and corrected if in error. The systems to be studied in this work use automatic-repeat-request (ARQ) to correct errors. If a message is detected to be in error, then the receiving node will request that the message be repeated. For the transmitting node to repeat the message, it must store all messages which it transmits until it is certain that they have been properly received. Thus the buffer contains both messages waiting to be transmitted and those waiting to be acknowledged. This study attempts to develop techniques which can minimize this holding time and total buffer occupancy and thereby increase the throughput between the two nodes.

Various methods have been used to return acknowledgements. In the early ARPA system, a short message carrying the acknowledgement was placed on the output buffer with priority over all other messages.[1] This method however degraded the message throughput because of the constant flow of acknowledgement messages. To alleviate the problem, a piggyback scheme was implemented. The purpose of this study is to compare two different piggyback schemes so that techniques can be developed which will minimize the holding time and thereby increase the throughput between the two nodes.

Two different types of piggyback return schemes will be compared. In both of these schemes, a certain number of bits of each message are reserved for the acknowledgement information of the reverse channel. In the first system, which we call pure piggyback (PP), when there are no messages to be transmitted, the acknowledgement information is stored until a message which can carry back this information is processed. In the second system, which we call ack generated piggyback (AGP), when a return message is not available, the acknowledgement information will not be stored. Instead, this information will generate its own acknowledgement message on which it can return to the originating node.

Previous work in this area has included that of A. Danthine and J. Bremer[2] and S. Lam.[3] Lam studied a model similar to our ack-generated piggyback system. However, he neglected the dependence that exists between the two nodes. Specifically, the length of time it takes for an acknowledgement to return to the originating node is dependent on the amount and type of traffic at the receiving node.

A queueing model which can take into account the effects of different acknowledgement protocols of a node pair is shown in Fig. 1 and is described by the following variables:

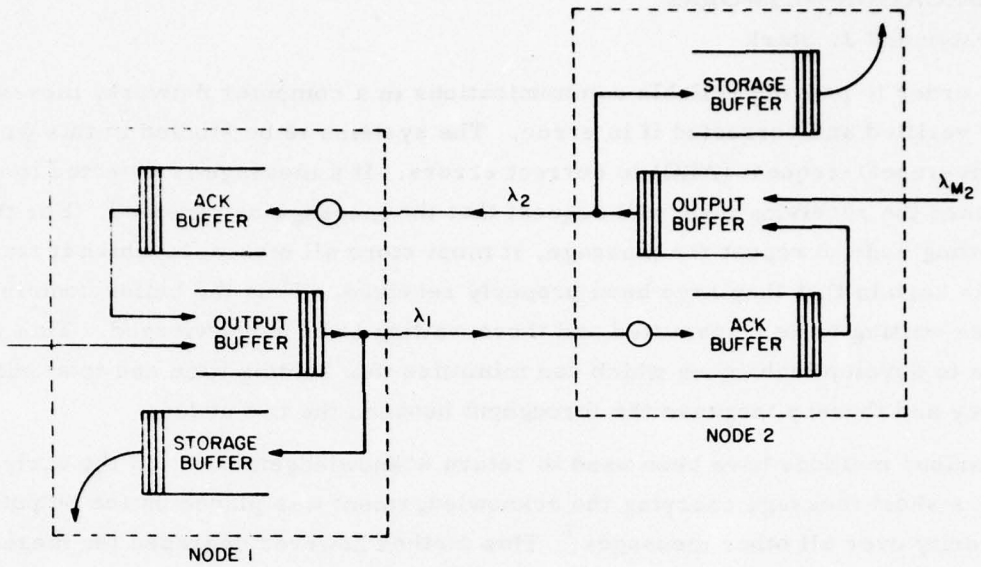


Fig. 1. Simplified queueing model of a node pair.

λ_{M1} = average number of messages per second which enter the originating node and are destined for the receiving node.

λ_1 = average total traffic which is transmitted from the originating node to the receiving node (messages per second).

A similar set of variables is also defined for the receiving node, where 1 and 2 are interchanged. The originating station is labeled 1 and the receiving station is labeled 2. For simplicity, errors in message and acknowledgement transmission and time-out errors will be ignored.

A. Pure Piggyback

In the pure piggyback system, each transmitted message has a certain number of bits which are reserved exclusively for the transmittal of all acknowledgements which are to be returned. The acknowledgement can only be transmitted as part of a message. Therefore, if there are no messages to be transmitted in one direction only, the acknowledgements of the messages in the opposite direction will queue, filling the available buffer space and cause all useful communication to cease.

In order to obtain initial solutions, certain assumptions were made in the model. It will be assumed that the error probabilities are zero. This implies that messages will rarely be repeated and therefore, the time-out property of the queue can be neglected. In addition, and for simplicity, it is assumed that all messages waiting to be acknowledged, will be properly acknowledged when the current message is received.

Once a message begins transmission its content is generally fixed for the duration. The acknowledgement information can be contained at the end of a message. Thus, if a message arrives while a return message is in the process of being transmitted, its acknowledgement could be added to the message. Although this is not usually done, we assume this operation here to simplify the model. These assumptions tend to reduce the blocking probability due to messages waiting to be acknowledged.

Due to the second assumption above, the number of messages waiting to be acknowledged at node 1 is equal to the number of acknowledgements in the ACK Queue of node 2. The probability of being in any state S is represented as $P_{M_1A_1M_2A_2}$ where

M_1 = number of messages to be transmitted at node 1
M_2 = number of messages to be transmitted at node 2
A_1 = number of messages that are waiting to be acknowledged at node 1
A_2 = number of messages that are waiting to be acknowledged at node 2

If N_1 and N_2 are the maximum buffer size at node 1 and node 2 respectively, then

$$N_1 \geq M_1 + A_1$$

and

$$N_2 \geq M_2 + A_2$$

We will first solve the problem for the case when the maximum buffer size is two, and then show how the problem would be solved for a general buffer size. Figure 2 shows the complete state diagram for the case when $N_1 = N_2 = 2$.

The set of states {(0000), (0010), (1000), (1010), (0020), (1020), (2020), (2010), (2000)} form an initialization set, which can be neglected when obtaining the steady state solution. The system can be in one of these states initially, but once the system leaves the initialization set, the system will never return to any state in this set. The system leaves the initialization state when the first message is transmitted.

Because of the previous assumption that an acknowledgement can piggyback on to messages which are in the process of being transmitted, the set of states {(0201), (0101), (0202), (0102)} are unreachable. Since the probability of being in any of these states when the system has reached steady state is zero, they can be neglected.

In order to simplify the solution somewhat, we will assume that the network is symmetric. Namely, that the input steady state traffic rate is the same for both nodes. Also the transmitting line speed is assumed identical in both directions between the two nodes. These assumptions are often valid for distributed networks. The assumption of equal traffic rates would usually not be correct, however, for centralized or

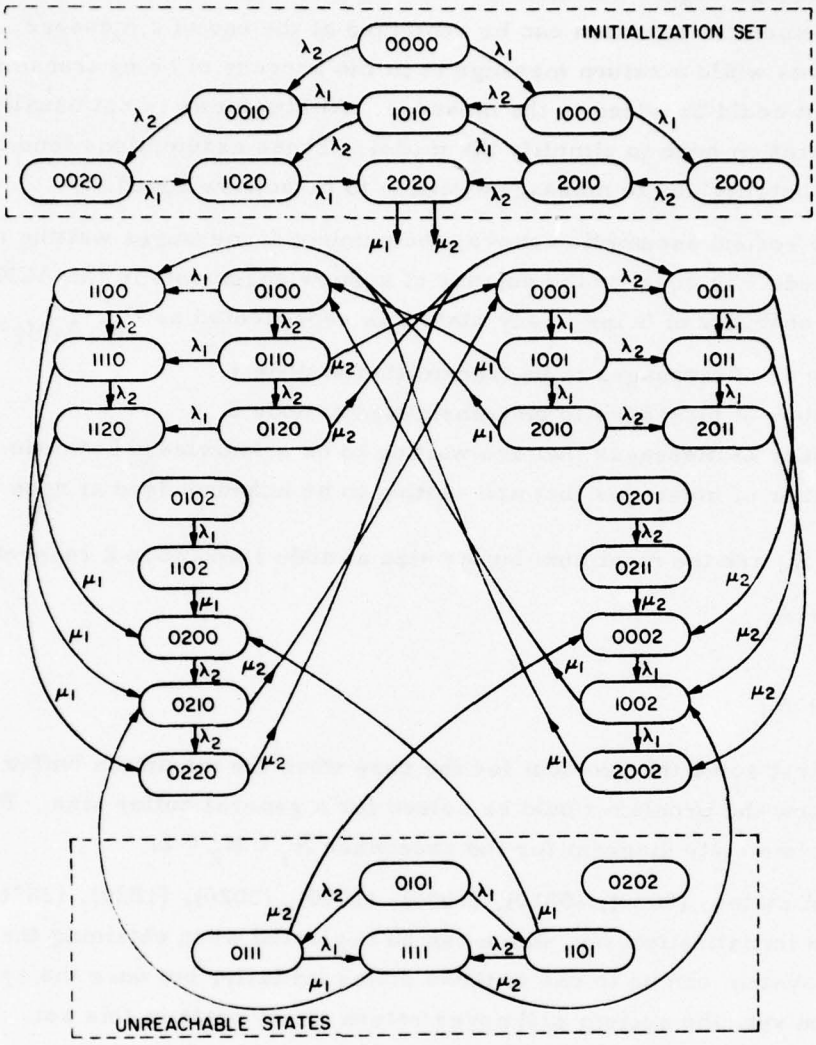


Fig. 2. Probability state diagram of the two buffer pure piggyback system.

star networks. It should be noted, that the acknowledgement system being discussed here should not be used in a network where the traffic rate is extremely unbalanced between the two nodes. In such a network, the transmitted messages of the busy node will have to wait an inordinate amount of time for acknowledgements to return, causing serious blocking problems at the node.

Due to the symmetric assumptions $P_{M_1A_1M_2A_2} = P_{M_2A_2M_1A_1}$ the probability state diagram for the two buffer case now reduces to nine states as shown in Figure 3. The set of state equation that need to be solved are easily found to be:

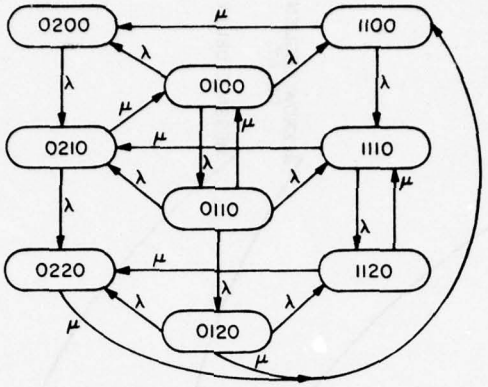


Fig. 3. Probability state diagram of the two buffer symmetric pure piggyback system.

$$2\rho P_{0100} = P_{0210} + P_{0110} \tag{1}$$

$$(\rho+1) P_{0110} = \rho P_{0100} + \rho P_{0120} + P_{1110} \tag{2}$$

$$P_{0120} = \rho P_{0110} \tag{3}$$

$$(\rho+1) P_{1100} = \rho P_{0100} + P_{0120} + P_{0220} \tag{4}$$

$$(\rho+2) P_{1110} = \rho P_{0110} + \rho P_{1100} + P_{1120} \tag{5}$$

$$P_{1120} = \rho P_{1110} + \rho P_{0120} \tag{6}$$

$$\rho P_{0200} = P_{1100} \tag{7}$$

$$(\rho+1) P_{0210} = \rho P_{0200} + P_{1110} \tag{8}$$

$$P_{0220} = \rho P_{0210} + P_{1120} \tag{9}$$

where

$$\rho = \frac{\lambda_1}{\mu_1 c_1} = \frac{\lambda_2}{\mu_2 c_2}$$

The solution of these equations and thereby the entire state diagram can be obtained numerically for specific values of ρ in the interval $0 \leq \rho \leq 1$.

Once the state probabilities of the system are known, then such system characteristics as the probability of blocking, the average time delay of message and scknowledgements, and the throughput of the network can be determined. The results have been plotted in Figures 4 to 8. There are three different means by which a node can be

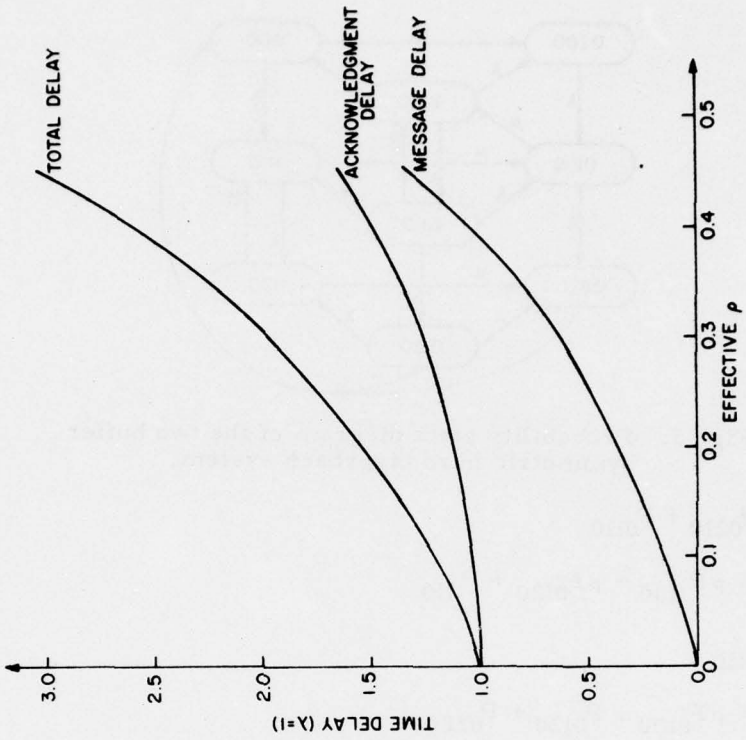


Fig. 5. Time delays of the two buffer, pure piggyback system.

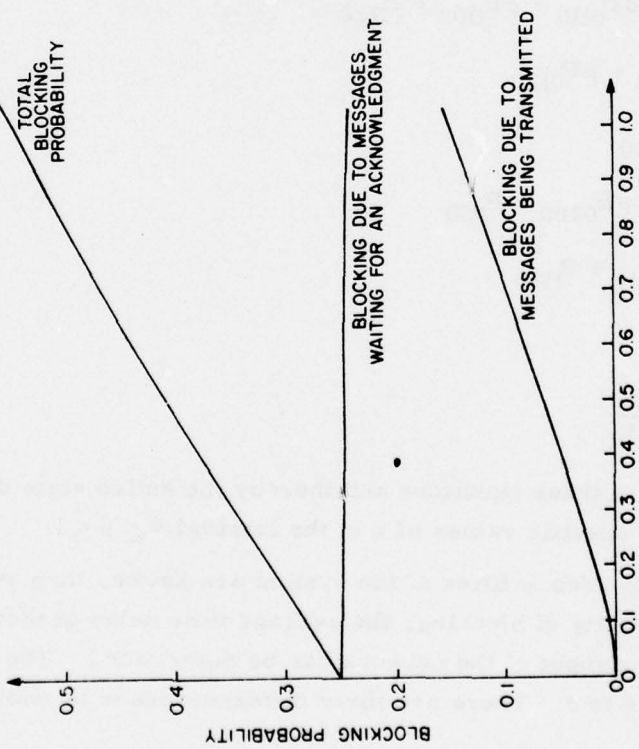


Fig. 4. Blocking probabilities of the two buffer, pure piggyback system.

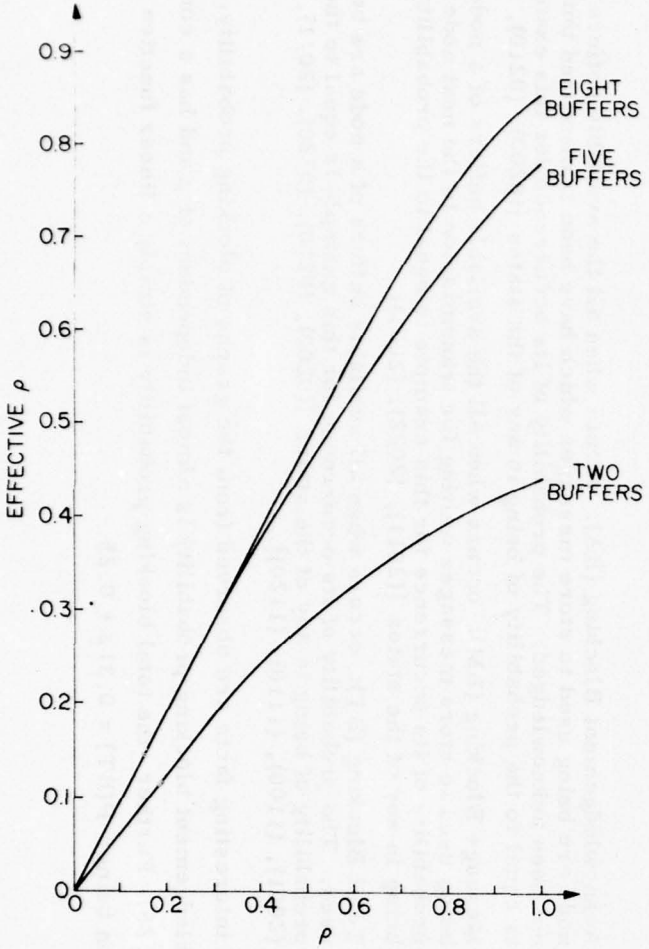


Fig. 6. Effective ρ for the pure piggyback system.

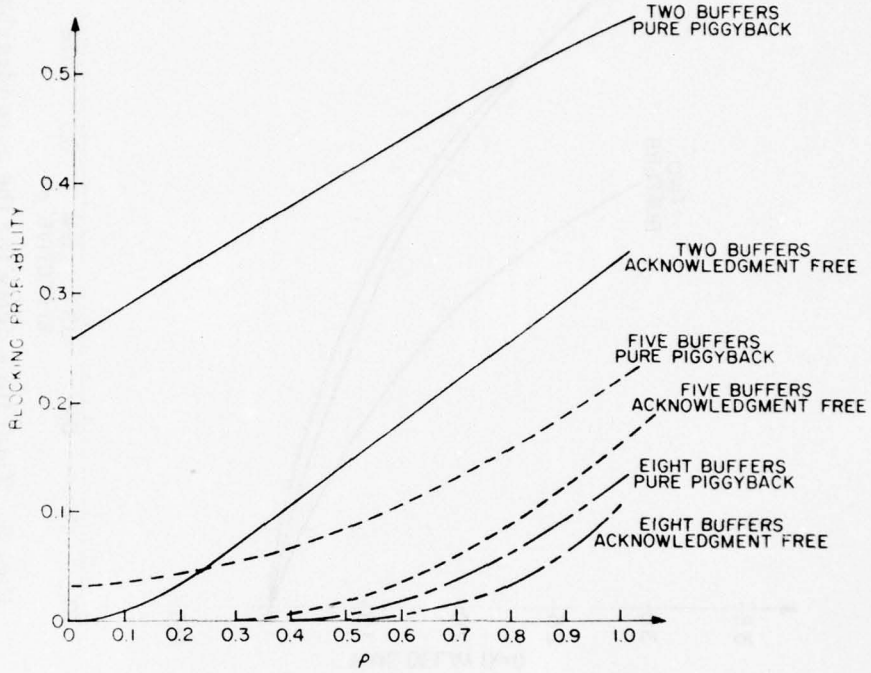


Fig. 7. Total blocking probabilities for the pure piggyback system and the acknowledgement free system.

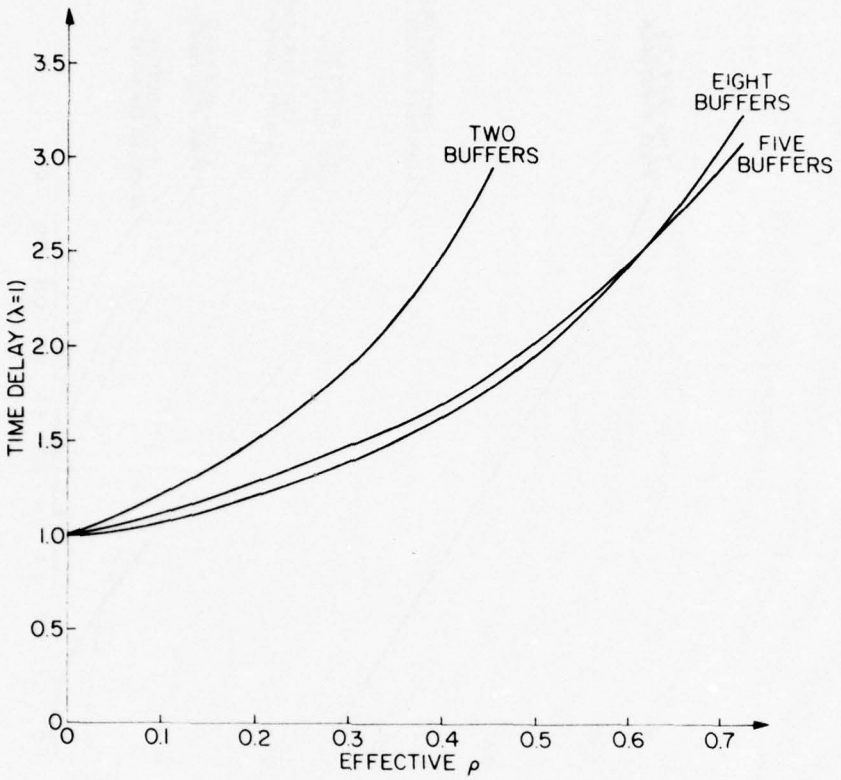


Fig. 8. Total time delay for the pure piggyback system.

blocked:

(1) Acknowledgement Blocking (BA): occurs when all the available buffers of a node are being used to store messages which have been transmitted but have not been acknowledged. The probability of its occurrence for this example is equal to the probability of being in any of the states {(0200), (0210), (0220)}.

(2) Message Blocking (BM): occurs when all the available buffers of a node are being used to store messages waiting for transmission to the next node. The probability of its occurrence for this example is equal to the probability of being in any of the states {(2011), (2022), (2001)}.

(3) Total Blocking (BT): occurs when all available buffers of a node are being used. The probability of its occurrence for this example is equal to the probability of being in any of the states {(0200), (0210), (0220), (2011), (2002), (2001), (1100), (1110), (1120)}.

Two interesting facts are observed from the graphs of blocking probability. First, the acknowledgement blocking probability is almost independent of ρ and has a constant value of 0.25. Further, the total blocking probability is simply a linear function of ρ, the function being: $P(BT) = 0.31\rho + 0.25$.

In order to compare the various cases which are to be studied, it is necessary to introduce the concept of effective ρ. It is determined by using the mean rate of messages that actually enter the system and are not blocked. This arrival rate is defined as

$$\lambda' = \lambda(1 - P(BT)) \quad .$$

The effective ρ is actually a measure of the throughput of the system. While it is possible to operate the system with the value of ρ greater than one, the value of effective ρ can never be greater than one. This however would not be a very efficient system since most messages would be blocked and have to be repeated. The effective ρ which is plotted in Fig. 6, starts off equal to ρ initially, but then begins to increase less rapidly and eventually would become constant for large values of ρ. This type of curve is true for all cases which we have studied.

The total delay time of a message is defined as the sum of the message delay time from the transmitting node to the receiving node plus the delay time of the returning acknowledgement. This is the total time that the buffer must store a message. The various time delays are found by application of Little's formula. The average number of messages in each of the queues is found from the computer results. This value is divided by the actual rate of messages entering the system. Since it is assumed that messages and acknowledgements are not lost in the system, the acknowledgement arrival rate is equal to the message arrival rate.

The message delay time initially starts at zero and then increases rapidly as the effective ρ increases. The total delay time, however, starts at one unit of time and then increases rapidly as a function of effective ρ. This peculiarity is due to the delay time of an acknowledgement in this system. Even with no new input traffic the previous message must still be acknowledged. However, the acknowledgement can not be transmitted until a new message arrives. This minimum delay time would be reduced if a time-out period of less than one time unit was used.

B. Solution for General Buffer Size

In the previous example the storage buffer was limited to only two storage locations. While this is an interesting problem, in most practical systems, buffer size would be larger. In order to obtain some insight into this problem, the Markov state diagram must first be obtained. This diagram for the general buffer case has been obtained and the appropriate set of difference equations developed.

To find the state probabilities from these equations is extremely complex. In order to solve these equations a computer program, capable of obtaining solutions for any finite buffer size, was written. To write the balance equation for every state very

quickly becomes an enormous and time consuming task for the computer. This is apparent when one realizes that the number of states in this case is equal to $N(N+1)^2/2$, where N is the maximum number of messages a buffer can store. We have used a technique which reduces the number of equations to N^2. The set of N^2 equations which must eventually be solved only contains the unknown states $P_{M_1A_1M_2A_2}$, where $\{A_1 = 1, A_2 = 0, N-1 \geq M_1 \geq 0, N-1 \geq M_2 \geq 0\}$. These are just the states of the first plane ($A_1 = 1$) of the state diagram. It is possible to obtain all the other states in terms of these first plane states. For example, the states of the first row ($M_1 \geq 0$, $M_2 = 0$, $A_1 > 1$) can all be written in terms of the states $\{P_{M_1100},\ n-1 \geq M_1 \geq 0\}$. Similarly, the state probabilities of the second row $\{P_{M_1A_110},\ N-A_1 > M_1 > 0,\ N > A_1 > 2\}$ can be written in terms of the state probabilities of the first two rows of the first plane $\{P_{M_1M_20},\ N-1 > M_1 \geq 0,\ 1 \geq M_2 > 0\}$. This process can be continued until all state probabilities have been found. By use of this algorithm we were able to obtain solutions for a maximum buffer size of 8, in less than 3 minutes and using only 180K bytes of storage.

C. Results for the General Buffer Case

Using the computer program described above, this example was solved for maximum buffer sizes of five and eight. In these cases, the effects of blocking are of less importance. Therefore, the value of effective ρ does not deviate as much from the actual value of ρ. The various blocking probabilities which were previously defined can now be determined from the following general equations:

$$P(BA) = \sum_{M_2=0}^{N} P_{0NM_20}$$

$$P(BM) = \sum_{M_1=0}^{N-1} \sum_{A_1=1}^{N-M_1} P_{M_1A_1N0}$$

$$P(BT) = P(BA) + P(BM) + \sum_{M_1=1}^{N-1} \sum_{M_2=0}^{N} P_{M_1N-M_1M_20}$$

Similarly, the various average time delays can be determined by applying Little's formula to the average number in each buffer, which are determined from the following equations:

$$\bar{N} = \sum_{M_2=0}^{N} \sum_{M_1=0}^{N-1} \sum_{A_1=1}^{M_1-N} A_1 P_{M_1A_1M_20}$$

$$\bar{N}_M = \sum_{M_2=0}^{N} \sum_{M_1=0}^{N-1} \sum_{A_1=1}^{M_1-N} (M_1 + M_2) P_{M_1A_1M_20}$$

$$\bar{N}_T = \bar{N}_A + \bar{N}_M$$

The effective message input rate is determined exactly as before

$$\lambda' = \lambda(1 - P(BT)).$$

The results for these two cases have been plotted in Figures 6, 7, and 8. As is to be expected, the blocking probabilities decrease dramatically as the buffer size is increased. From Fig. 7 it is seen that the blocking probability curve of the 8 buffer system can be approximated by that of an acknowledgement free system. It should be noted that the effective ρ, which is related to the blocking probability, for the 8 buffer case is approximately equal to ρ until the traffic rate becomes very high. This of course is to be expected, since so few incoming messages must be rejected by the system for this case.

The acknowledgement time delay again has an initial value of one, but now remains virtually constant for all values of ρ. Of greater interest is the total time that a message is stored at a node. This characteristic is interesting in that when it is plotted against ρ, the average time delay for large ρ increases as the buffer size increases. This is not surprising, though when one realizes that due to blocking the small buffer limits the maximum message delay. For the larger buffer, however, these are occasions, although rare when the maximum of eight messages are awaiting transmission. When plotted against effective ρ then it is seen that the total delay time of the 2 buffer case is indeed the poorest. Also, while the total time delay of the 8 buffer case is initially slightly lower than that of the 5 buffer case, it is increasing at a faster rate and finally surpasses that of the 5 buffer case when the effective ρ is extremely high.

From these results, it is apparent that while 2 buffers is not enough storage, 5 buffers should be sufficient. Increasing the number of buffers to 8, will only increase the throughput slightly and will not significantly improve the average time delay of the system.

D. Ack Generated Piggyback System

While the pure-piggyback system is quite good and easy to implement, it suffers from one major drawback. This is the problem of large time delays for acknowledgements under light load conditions. In practice, this problem is alleviated somewhat by the use of a time out mechanism. This is not a very satisfying solution since it will cause unnecessary repetition of many good messages. A far better system, which is to be studied here, is obtained through the use of ack generated messages. In this system, when it is required to return an acknowledgement but there are no messages being transmitted to the originating node, then a special ack message will be transmitted.

As in the case of regular messages, it is possible for this message to return more than one acknowledgement. However, the ack message need not have the same average size as a regular message. In our model, once the decision is made to create an ack message, it will be created and transmitted even if a regular message is received destined for the originating node. The ack messages are of course not to be acknowledged.

In order to solve the problem, the probability state diagram and its associate set of difference equations must first be determined. From this a computer program, similar to that used for the PP system, can be solved for the complete set of state probabilities of the AGP system. Since acknowledgements are not to be acknowledged, it is necessary to differentiate between an ack message and a regular message. This is accomplsihed by means of an additional dimension for the state probabilities. This of course, further complicates the state diagram and any possible solution. The state probabilities are now represented by, $P_{M_1 A_1 M_2 A_2 L}$ where,

$$L = \begin{cases} 0, \text{ if } M_2 \text{ is a regular message} \\ 1, \text{ if } M_2 \text{ is an acknowledgement message} \end{cases}$$

M_1, A_1, M_2, A_2 are the same as previously defined.

In order to allow the size of an acknowledgement message to be different from that of a regular message, another parameter s is used. This parameter s is defined as the ratio of the acknowledgement message size to that of the regular message size. As for the PP system, this system will first be solved for a maximum buffer size of two. It will then be shown how the system can be solved in the general buffer case.

It should be noted that the previous initialization state set, in which both acknowledgement buffers are empty, can now exist. However, the set of states for which an acknowledgement was waiting for a message does not now exist. The state diagram for this system does not possess the complete symmetry found in the PP system. There is no symmetric pair in the complete state diagram for the states (00000), (10100), (20200).

Again, the solution of the state equations can readily be obtained for specific values of ρ in the interval $0 < \rho \leq 1$. From these results, all the state probabilities can be determined and the entire system characterized.

To compare the two systems, similar system characteristics were determined for the AGP system as were determined for the previous PP system. A plot of P(BT) vs. ρ for various values of the parameter s is shown in Figure 9. For this system P(BT) is initially zero. Of interest is the fact that P(BT) for this system, when the

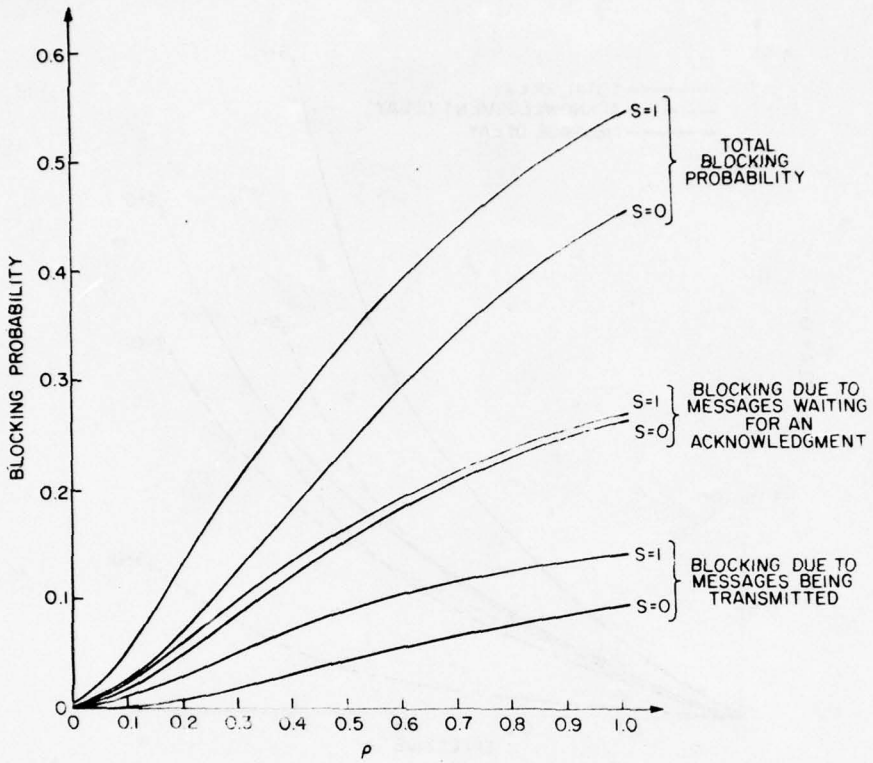


Fig. 9. Blocking probabilities of the two buffer ack generated piggyback system.

acknowledgement message size is equal to the message size, approaches asymptotically the curve of P(BT) for the PP system. This of course, is to be expected. When ρ is close to one, there will be a continuous stream of messages to be transmitted, so that there will always be a message available on which an acknowledgement can be piggybacked. It should also be noted that as the acknowledgement size is decreased to zero, the blocking probability is improved by at least 20% for large values of ρ, and by over 50% for small values of ρ. This improvement is due to the decrease of the average delay time of an acknowledgement, especially for low values of ρ.

The various time delays were also plotted against effective ρ in Figure 10. These plots show the tremendous improvement possible, as the acknowledgement size is decreased. When $s > .5$, the average message delay time of this system is greater than that of the PP system. This is due to the increase in traffic because of the acknowledgement messages that are transmitted. However, this is not the case for the more important characteristics of total average delay. For this system the total time delay is initially zero and then increases approaching the total time delay of the PP system for

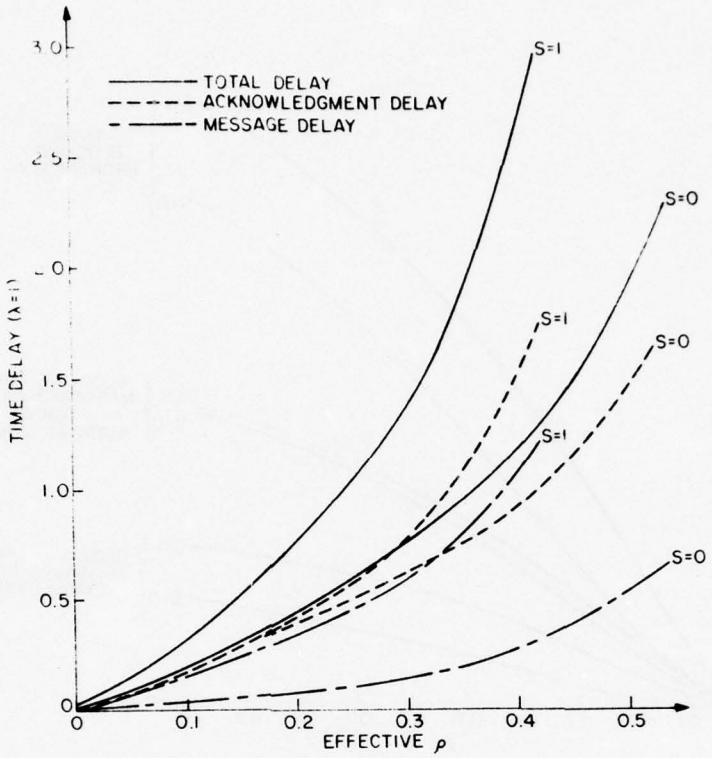


Fig. 10. Time delays of the two buffer ack generated piggyback system.

large values of ρ. In the case of s = 1, the two curves actually cross when $\rho' = .4$. Above this value the time delay of the PP system is better than that of the AGP system.

E. Solution for General Buffer Size

As was done for the Pure Piggyback system, this system was solved for a general buffer size. This problem is solved by a similar type of computer program as was used for the pure piggyback system. The program reduces the number of unique states from $\frac{1}{2}(N+1)(2N^3+N+2)$ to $\frac{1}{2}(3N^2+3N+2)$, where N is the maximum number of messages a buffer can store. The set of states that are to be solved are:

$$P_{M_1 0 M_2 00}, \begin{cases} 0 \le M_1 \le N \\ M_1 \le M_2 \le N \end{cases}$$

and

$$P_{M_1 1 M_2 00}, \begin{cases} 0 \le M_1 \le N-1 \\ 1 \le M_2 \le N \end{cases}$$

Once these state probabilities are determined, the program can solve for all the state probabilities of the system. By use of the computer algorithm, a solution for a maximum buffer size of 5 was obtained.

F. Results for the General Buffer Case

By means of the computer program, the various system characteristics as a function of ρ were obtained. The set of blocking probability curves, shown in Fig, 12, are similar to blocking probability curve of the 5 buffer pure piggyback system. Again, the AGP system is better at low values of ρ, but is similar to the PP system at large values of ρ. The effect of the acknowledgement message size is not as great as it was for the 2 buffer case, since the blocking probability is already low. Similar results are obtained for the various time delays. The message time delay for the AGP system is greater than for the PP system. However, the total delay for the AGP system is less than that of the PP system, for values of $\rho' \leq .65$ for $s = 1$, and $\rho' \leq .77$ for $s = 0$. Above these values, the PP system is better. The value of effective ρ, plotted in Fig. 11,

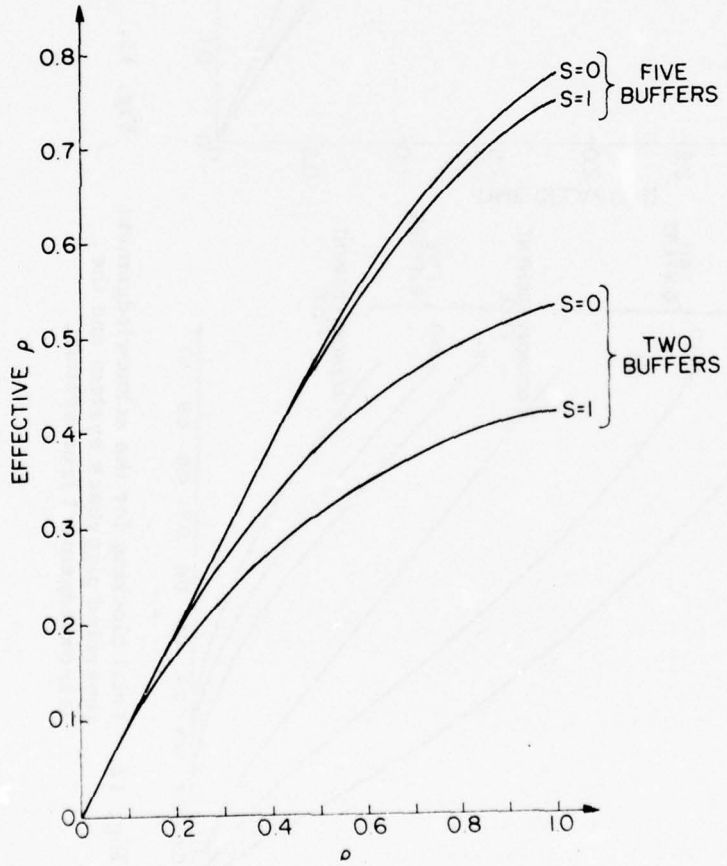


Fig. 11. Effective ρ for the ack generated piggyback system.

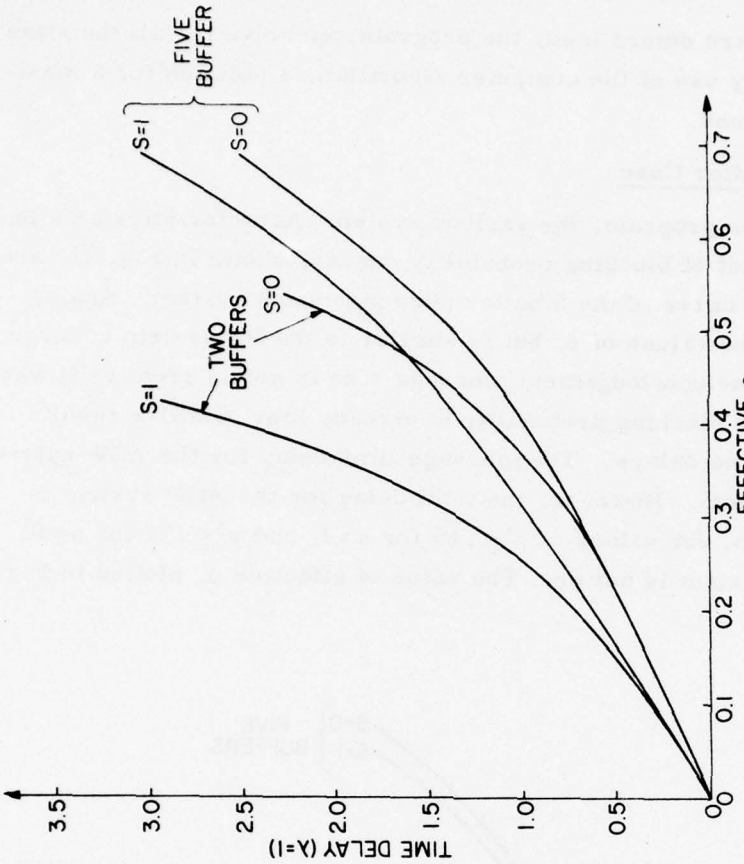


Fig. 13. Total time delay for the Ack generated piggyback system.

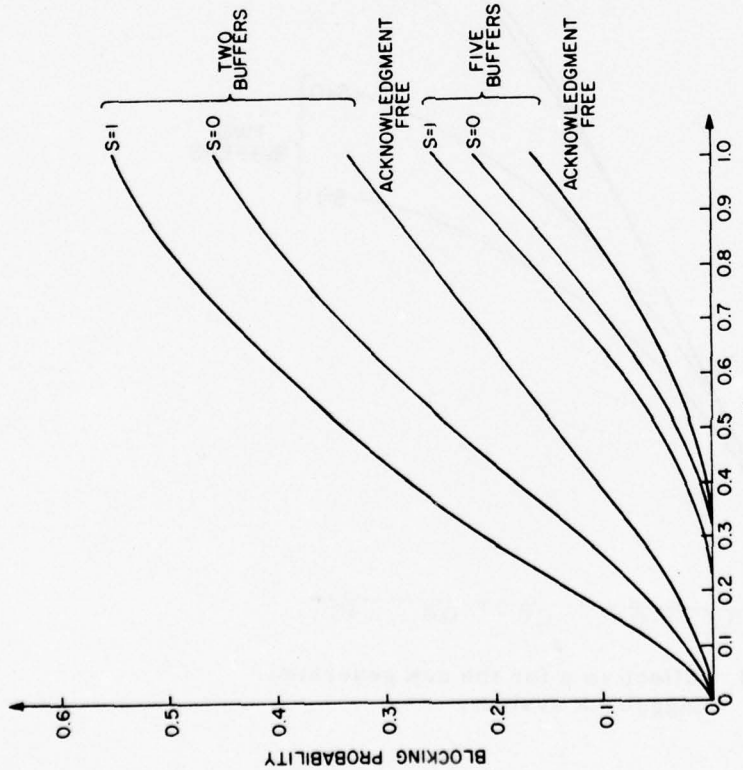


Fig. 12. Total blocking for the acknowledgement generated piggyback system and the acknowledgement free system.

is not significantly affected by the acknowledgement message size. It is approximately equal to ρ, until $\rho > .6$. The effective ρ for the AGP system is slightly better than that of the PP system, until $\rho = .75$ for $s = 1$, and until $\rho = .9$ for $s = 0$. Above these values the effective ρ of the PP system is larger. It should be remembered that the effective ρ is proportional to the throughput of the system. Figure 13 gives the time delay under similar conditions.

G. Conclusions and Future Work

From the results of the work reported here, it appears that 5 buffers would usually be sufficient. The AGP system is better than the PP system at low to moderate traffic loads. At heavy traffic load, the reverse is true. This is due to the decrease in overhead required by the PP system. The point where this reversal occurs is primarily dependent on the acknowledgement message size of the AGP system. This reversal even occurs when the size of the acknowledgement size is zero. These results are interesting, since a PP system would usually be easier to implement. If the traffic load had a large variance, than an adaptive system, switching between the two acknowledgement schemes would be beneficial.

It is still necessary to study the two systems when some of the assumptions which were used are removed. Primarily this includes the lack of a time-out period and the requirement of a symmetrical network. Such analysis is planned for the future.

Joint Services Technical Advisory Committee
F44620-74-C-0056

R. Boorstyn and J. Stark

National Science Foundation
ENG 73-04232

REFERENCES

1. J.N. McQuillan, D.C. Walden, F.E. Heart, B.P. Cosell and W.R. Crowther, "Improvements in the Design and Performance of the ARPA Network," Fall Joint Computer Conference, AFIPS Conference Proceedings, 41, 741-754 (1972).
2. A. Danthine and J. Bremer, "Packet Node Behavior of the Internode Protocol," IEEE Trans. Communications, COM-24, 606-614 (June 1976).
3. S.S. Lam, "Store and Forward Buffer Requirements in a Packet Switching Network," IEEE Trans. Communications, COM-24, 394-403 (April 1976).

EFFECTS OF MANPOWER DEPLOYMENT AND ERROR GENERATION ON SOFTWARE RELIABILITY

M. Shooman and S. Natarajan

Even though the software field is nearly two decades old it is only now that the attention is focused on the understanding of the important aspects of bug generation and removal inherent in most computer programs. In this contribution we shall discuss the dynamics of error generation and error removal during the debugging process.[1]

Earlier work[2] assumed that the total numbers of errors (i.e., the sum of those corrected and those remaining) in a program remained constant throughout the debugging process. This leads to the conclusion that if the debugging continues long enough we will eventually remove all the errors. In practice the contrary is true. Firstly, we create new errors during debugging, and secondly we stop the debugging process before we have removed all errors. Thus the total number of errors does not remain constant.

We will list some of the ways in which errors are generated:

(1) The correction of a bug may work locally only (i.e., the global aspects of the error still remain)

(2) A typographical error may arise as the result of bug correction

(3) The correction is based upon faulty analysis and does not accomplish any bug removal.

(4) The correction is accomplished, however, it is accompanied by the creation of a new error.

A. Experimental Debugging Data

Prior to the formulation of an error model we shall investigate the available error data. Shooman's earlier work presents error correction data from seven different programs. These are portrayed in Figures 1 and 2. Cumulative error correction curves are shown in Figures 3 and 4.

Shooman and Bolsky[3] report the results of a software debugging experiment. One of the outcomes of this experiment is shown in Figure 5. We observe from the figure that the time expended on fixing the earlier and the later discovered bugs is the same. This leads us to postulate that the rate of correction of errors is constant, counteracting the general belief that the later bugs are harder to correct. Subsequently we shall show that the use of the constant removal rate hypothesis in postulating a model leads to anomalous results. (See Ref. 1 for a discussion of the validity of the data in Figure 5.)

Another study on software debugging was conducted by Akiyama.[4] His analysis is shown in Figure 6. If we examine Figs. 3, 4 and 6 we find the cumulative correction

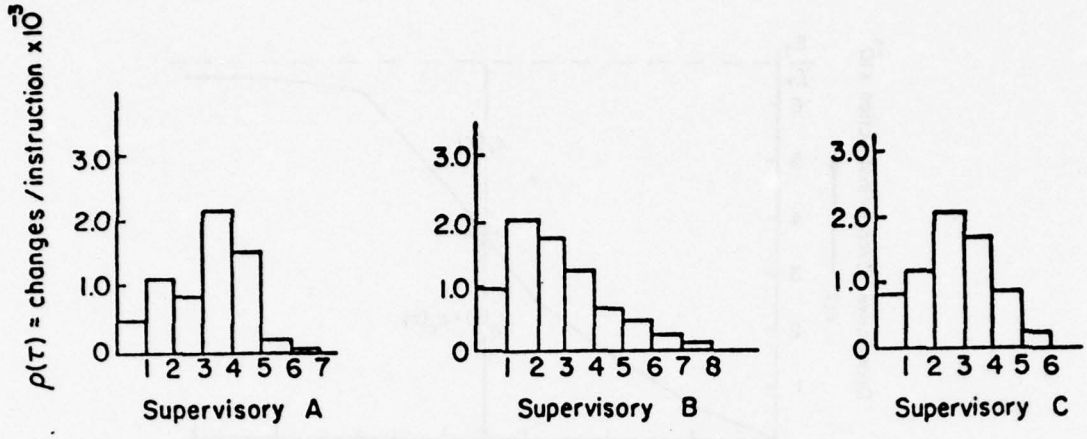


Fig. 1. Normalized error rate vs. debugging time for three supervisory programs

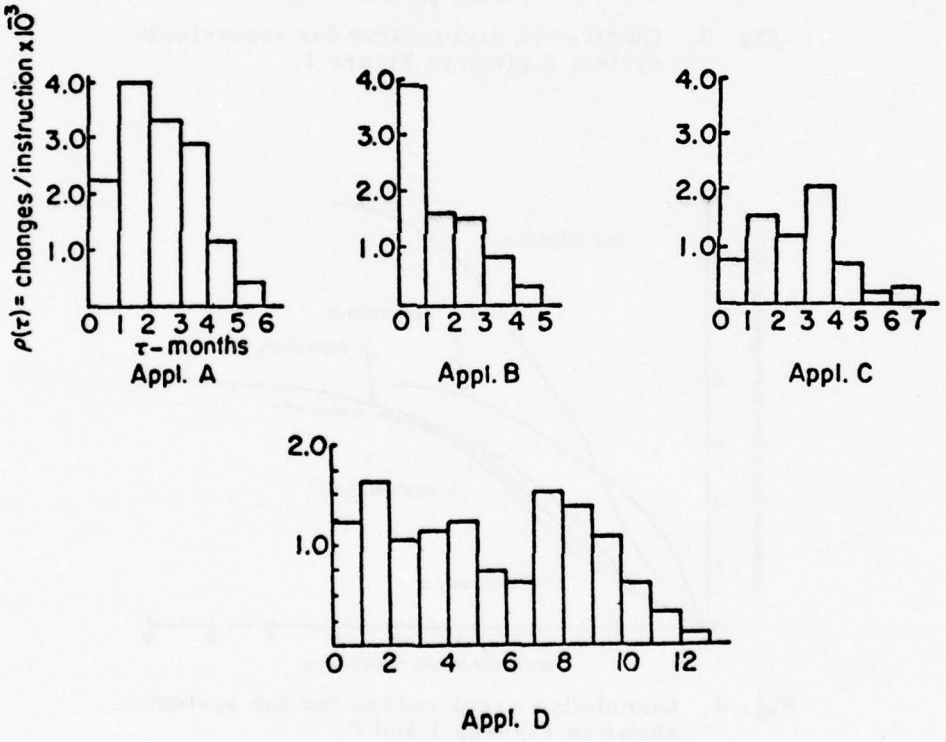


Fig. 2. Normalized error rate vs. debugging time for four applications programs.

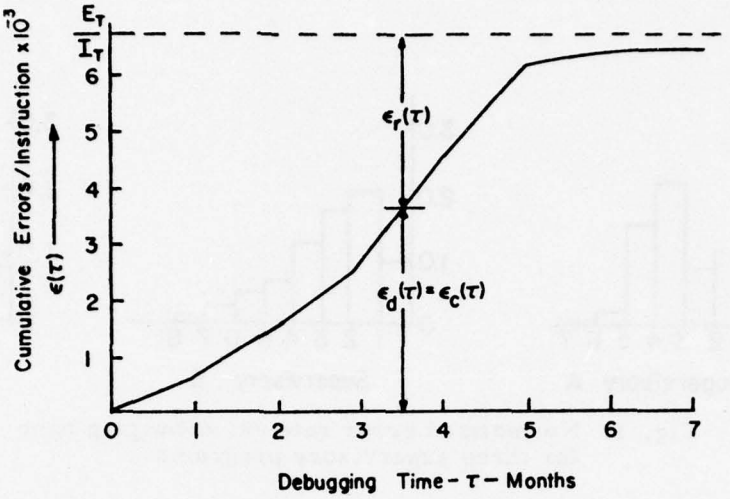


Fig. 3. Cumulative error curve for supervisory system A given in Figure 1.

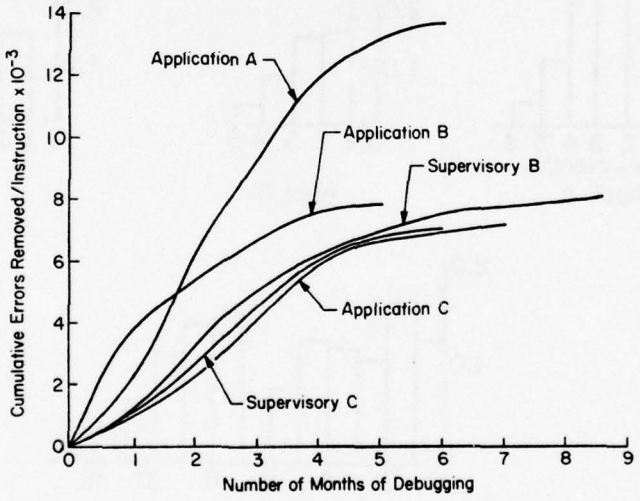


Fig. 4. Cumulative error curves for the systems shown in Figures 1 and 2.

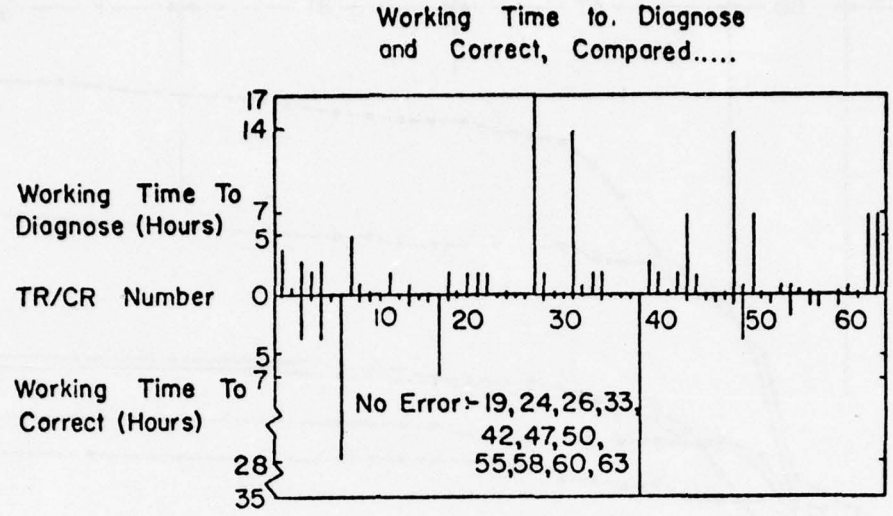


Fig. 5(a). Working time expended in debugging as observed at Bell Labs.

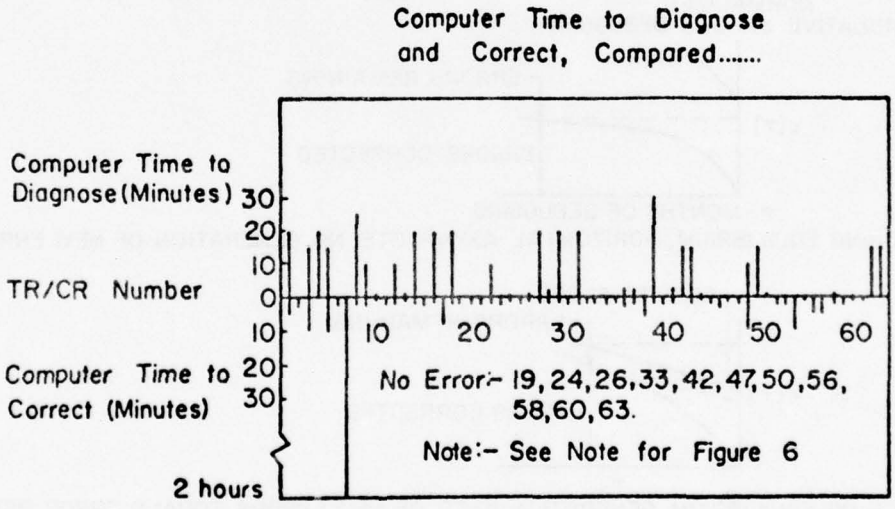


Fig. 5(b). Computer time expended in debugging as observed at Bell Labs.

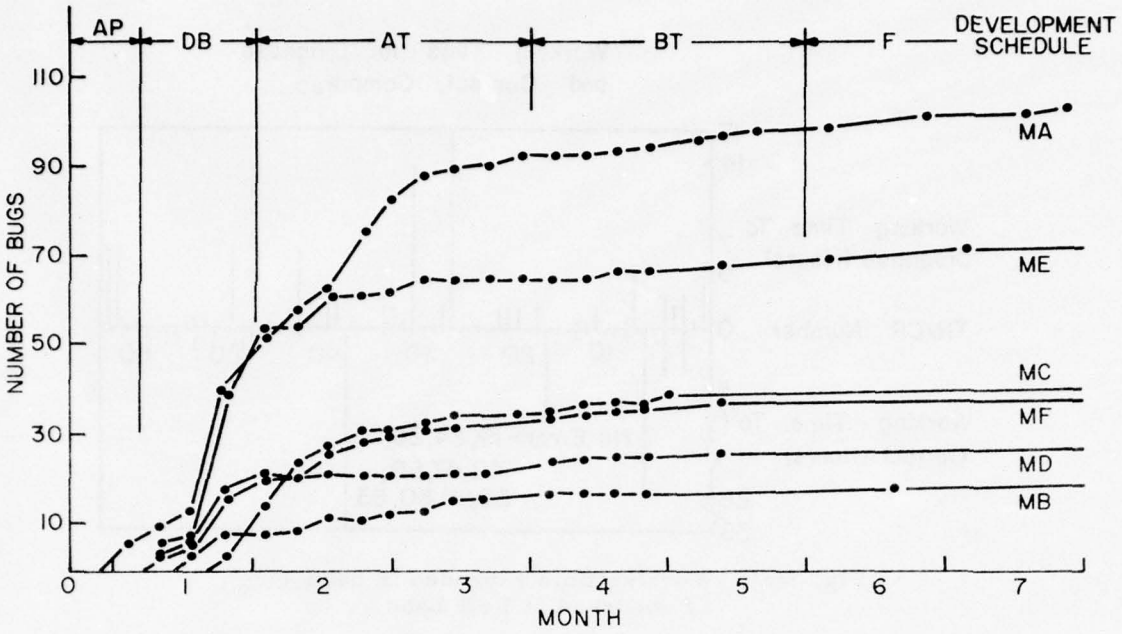
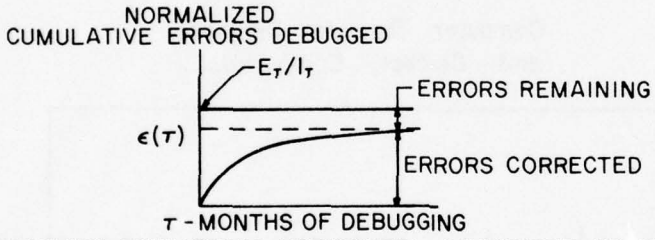
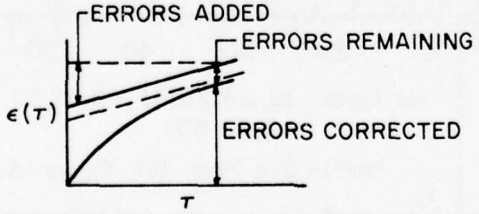


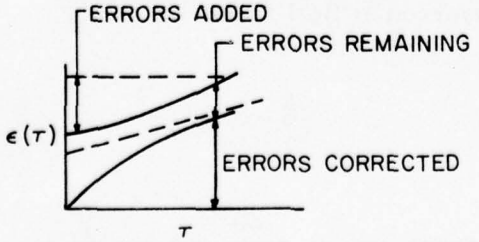
Fig. 6. Cumulative curves on the occurrence of bugs for each module of sample.



(a) APPROACHING EQUILIBRIUM, HORIZONTAL ASYMPTOTE, NO GENERATION OF NEW ERRORS.



(b) APPROACHING EQUILIBRIUM, GENERATION RATE OF NEW ERRORS EQUALS ERROR REMOVAL RATE.



(c) DIVERGING PROCESS, GENERATION RATE OF NEW ERRORS EXCEEDS ERROR REMOVAL RATE.

Fig. 7. Cumulative errors debugged vs. months of debugging.

curve has a high initial slope and gradually decreases as debugging proceeds. We will make use of this fact in our hypothetical model.

B. The Removal-Generation Error Model

We shall consider three cases of software debugging:

(1) Error generation rate < Error correction rate

(2) Error generation rate = Error correction rate

(3) Error generation rate > Error correction rate.

These three cases are shown in Figure 7.

Intuitively we may feel that bug generation rate $r_g(\tau)$ is proportional to the number of errors $n(\tau)$ remaining in the program. If we introduce the detection rate, $r_d(\tau)$ and if we further assume that not all detected errors are corrected we obtain Eqs. (1) and (2) for the correction rate $r_c(\tau)$ and the generation rate $r_g(\tau)$.

$$r_g(\tau) = a\, n(\tau)\, r_d(\tau) \tag{1}$$

where a is the proportionality constant.

$$r_c(\tau) = b\, r_d(\tau) \tag{2}$$

where b is the proportionality constant.

The rate of bug accumulation is given by Eq. (3)

$$\frac{d\, n(\tau)}{d\tau} = r_g(\tau) - r_c(\tau) \tag{3}$$

If the detection rate is constant, that is $r_d(\tau) = r_o$ where r_o is a constant then

$$\frac{d\, n(\tau)}{d\tau} = a\, n(\tau)\, r_o - b\, r_o \tag{4}$$

The solution of Eq. (4) is

$$n(\tau) = (n_o - b/a)\, e^{a r_o \tau} + b/a \tag{5}$$

where n_o is the initial number of errors at $\tau = 0$. Equation (5) gives rise to three cases which are given below.

(1) We discount the case where $n_o = b/a$ because the probability of such an occurrence is very low.

(2) If $n_o > b/a$ the errors build up with resulting instability

(3) If $n_o < b/a$ then after prolonged debugging we find from Eq. (5) that $n(\tau)$ decreases and becomes negative which is an absurd result.* In summary the following two conclusions may be drawn.

---

*The number of errors can not be a negative quantity.

(a) Either the assumptions are wrong, or

(b) The model is valid for only a very short period.

The result is depicted in Figure 8.

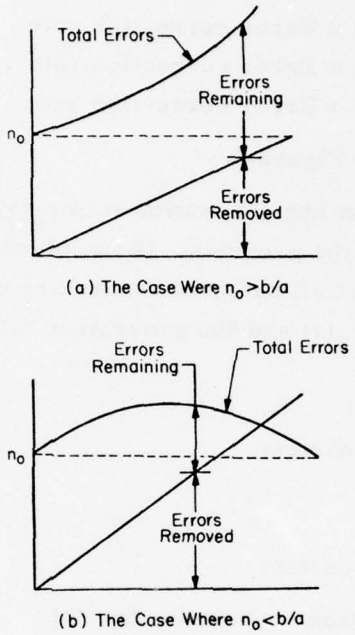


Fig. 8. The model developed with the assumption that the correction rate is a constant, while the generation rate is proportional to the number of remaining errors.

If we let a $r_d(\tau) = a_1$, then Eq. (1) reduces to

$$r_g(\tau) = a_1 n(\tau) \tag{6}$$

Let us assume that the error correction process follows two different laws. These two laws are given in Equations (7a) and (7b).

$$r_c(\tau) = \begin{cases} b_1 & \text{for} \quad n(\tau) > n_1 \quad (7a) \\ b_2 n(\tau) & \text{for} \quad n(\tau) \leq n_1 \quad (7b) \end{cases}$$

where r_1 is some critical number of bugs. If we substitute Eqs. (6) and (7a) in Eq. (3) we get

$$\frac{d n(\tau)}{d\tau} = a_1 n - b_1 \quad \text{for} \quad n(\tau) > n_1 \tag{8}$$

Solving Eq. (8) we get Eq. (9)

$$n(\tau) = (n_o - b_1/a_1)e^{a_1\tau} + b_1/a_1$$

once again by substituting Eqs. (6) and (7b) in Eq. (3) we get

$$\frac{d\,n(\tau)}{d\tau} = a_1 n - b_2 n \quad \text{for} \quad n(\tau) \leq n_1$$

A solution of Eq. (10) is

$$n(\tau) = n_1 \exp[(a_1 - b_2)(\tau - \tau_1)]$$

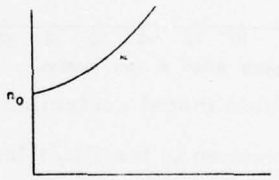
Equations (9) and (11) describe the error behavior throughout the debugging
We can differentiate among three cases depending upon the relative values c
meters. These are

Case 1. $n_o > b_1/a_1$ resulting in instability

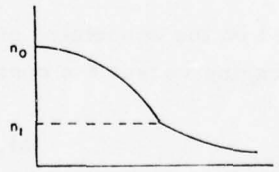
Case 2. $n_o < b_1/a_1$ and $a_1 < b_2$ depicting efficient debugging

Case 3. $n_o < b_1/a_1$ but $a_1 > b_2$ leading to oscillations.

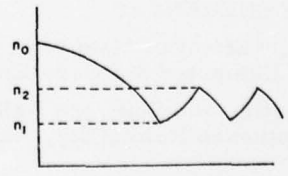
For Case 3, because $a_1 > b_2$, the errors build up again thereby forcir
rection rate to obey Equation (7a). Subsequently the error behavior is desc
Equation (9). These three cases are shown qualitatively in Figure 9.



(a) Unstable Model. Errors build-up indiscriminately.



(b) Controlled Model. Debugging is efficient.



(c) Oscillatory Model.

Fig. 9. Remaining errors plotted as a function of man-months of debugging.

In Case 1 the debugging process results in an error buildup. Software managers recognize such a problem, occurring unfortunately on many projects. The cure is to bring in more personnel or change the software team. Let us assume that this occurs when $n = n'$ at time $\tau = \tau'$. The new team may establish new values of a_1' and b_1' for a and b, such that $n' < b_1'/a_1'$. The number of errors now starts decaying. The three cases are shown in Figures 10, 11 and 12.

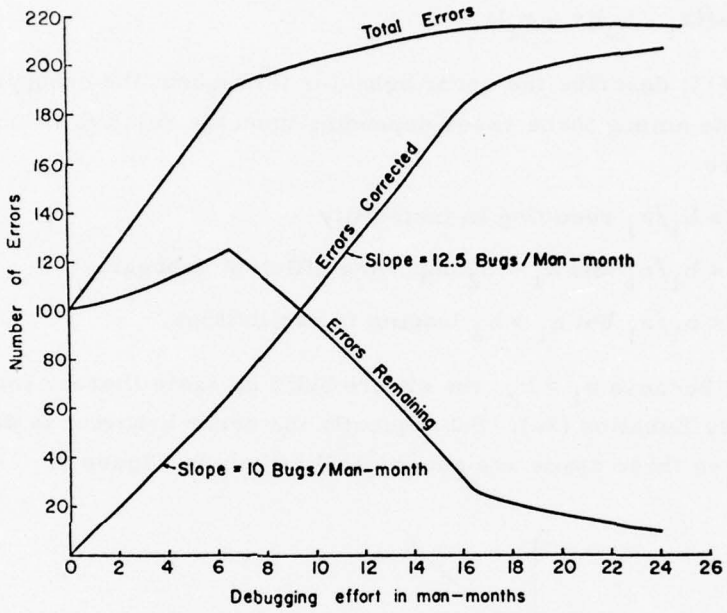


Fig. 10. Case 1. Unstable model controlled subsequently.

A computer program has been written in the PL/1 language to plot these curves for various data. The program plots the remaining errors, the cumulative corrected errors and the total errors.

Further study is being conducted on the economics of debugging for Case 1. The analysis will contrast the cost of debugging versus the cost of rewriting the program.

Rome Air Development Center
F-30602-74-C-0294

M. Shooman and S. Natarajan

Joint Services Technical Advisory Committee
F44620-74-C-0056

REFERENCES

1. M. L. Shooman and S. Natarajan, "Effect of Manpower Deployment and Bug Generation," Proc. MRI Symposium on Computer Software Engineering (April 20-22, 1976).
2. M. L. Shooman, "Operational Testing and Software Reliability Estimation," 1973 IEEE Symposium on Computer Software Reliability, New York City (April 30-May 2, 1973).
3. M. L. Shooman and M.I. Bolsky, "Types, Distributions and Test Correction Times for Programming Errors," Proc. 1975 Intl. Conf. on Reliable Software, Los Angeles (April 21-23, 1975).
4. F. Akiyama, "An Example of Software System Debugging," IFIP Congress 71, Ljubljana, Yugoslavia (August 1971).

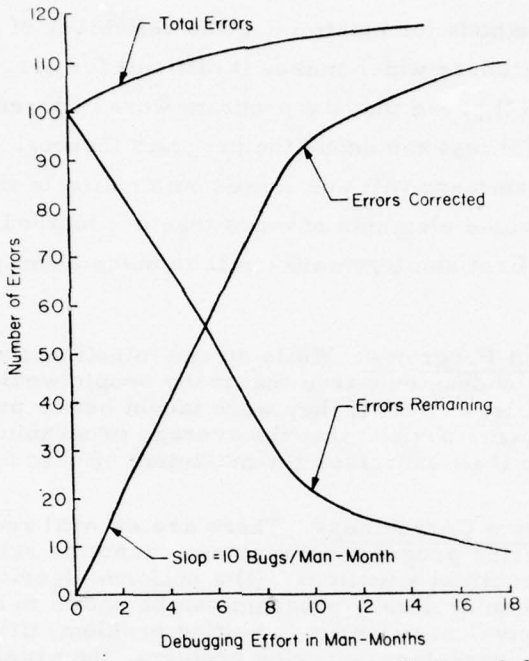


Fig. 11. Case 2. The controlled model.

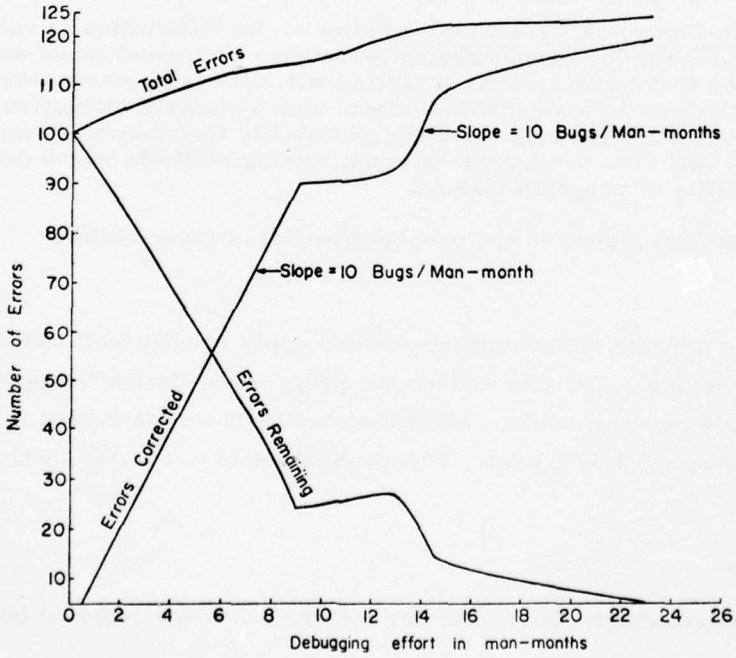


Fig. 12. Case 3. Oscillatory model. Oscillations controlled after the first bump.

STATISTICAL THEORY OF PROGRAM TESTING

A. E. Laemmel

There are three methods for maximizing the reliability of computer programs: (1) use a systematic procedure which makes it difficult for errors to occur during the writing of the program, (2) prove that the program works correctly by some formal or automatic process, and (3) test and debug the program thoroughly before passing it on to the user. Most programmers will use some combination of these methods, and in fact some procedures involve elements of more than one method. The present report emphasizes testing, but first some remarks will be made about program writing and proving.

(1) Writing Correct Programs: While no one intentionally inserts errors in his program, it is undoubtedly true that many people would produce more reliable programs with less effort if they were taught better programming techniques. However, it seems obvious that the average programmer should test his programs even if he exercises the maximum of care and uses the best techniques.

(2) Proving Program Correctness: There are several reasons why a formal procedure for proving program correctness cannot be relied on to insure absence of errors in practical situations: (i) a uniform algorithm for proving the correctness of an arbitrary program can be shown to be impossible, being essentially equivalent to Turing's halting problem; (ii) even for solvable subclasses of the correctness-proving problem, the usual method (some improvement on Herbrand search) is so time consuming as to be impractical; (iii) there is always a possibility of error in the proving program, or in applying it to the program being tested.

(3) Testing Computer Programs: In view of the difficulties of validating a computer program by programming techniques or formal proof methods, it is believed that some amount of testing will always be necessary. The purpose of this report is to describe a model which shows the relationship between errors of different types and the probability that they will cause a program to fail, and also to suggest optimum testing methods which minimize the probability of program failure.

A fairly detailed sketch of the proposed model is given below.

A. Definitions

Some basic aspects of the testing process apply equally well to a computer program or to a physical device. For this reason the program or device being tested will be referred to simply as the module. After the module has constructed it is checked by a tester and then employed by a user. The probability of error, P_e, which occurs during use is given by

$$P_e = P_m P_u \tag{1}$$

where P_m is the probability that the tester misses all of the residual bugs in the module,

and P_u is the probability that the user then encounters one of the overlooked bugs. If exhaustive testing is possible the $P_m = 0$, since it is assumed that if a bug is found it is corrected and the whole process is started again. In most cases of interest exhaustive testing is not possible or practical. If there are no bugs then all of the probabilities are zero and this possibility appears as a special case in the analysis to follow.

B. Elementary Model

A simply case might be the following: The module has N possible input values, and each of these are equally likely to be chosen by the tester or by the user. Of these input values, W cause improper functioning of the module, but neither the tester nor the user knows which inputs cause errors or even how large W is. The tester chooses t inputs at random without keeping a record of inputs previously tested, i.e., sampling with replacement. Under these circumstances, $P_u = W/N$, $P_m = (1 - W/N)^t$ and

$$P_e = \frac{W}{N}\left(1 - \frac{W}{N}\right)^t \leq \frac{W}{N} e^{-Wt/N} \tag{2}$$

A plot of P_e vs. W is displayed in Fig. 1. If the testing is to do any good, i.e., to reduce P_e to significantly less than W/N, then it is necessary that

$$t >> \frac{N}{W} \tag{3}$$

In many applications it is found that satisfying the inequality of Eq. (3) requires a very large number of tests t, and that our intuitive feeling is that P_e is acceptably small in spite of the testing using far fewer tests than indicated.

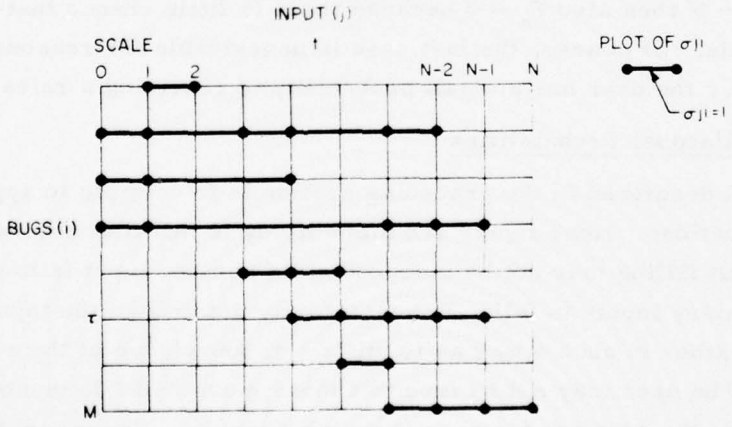


Fig. 1. Graphical representation of relationships among bugs and tests.

It will continue to be assumed that the tester passes no information to the user about which inputs were tested. It will also be assumed that each test either succeeds or fails; there is no additional information which would permit sequential sampling methods. Sampling without replacement would slightly lower P_m to

$$P_m = \begin{cases} \frac{(N-W)!\,(N-t)!}{(N-W-t)!\,N!} & t \leq N-W \\ 0 & t > N-W \end{cases} \qquad (4)$$

Note that this method requires that the tester keep a record of inputs already tested, or that he avoids duplication by other means.

The particular type of error to which P_e pertains must be borne in mind to avoid confusion with other possible definitions of error. If $W = N$ all inputs to the module cause malfunctions but $P_e = 0$ according to Equations (2) and (4). This is so, because the tester removes user bugs with each test, and continues until all tests are exhausted. In this case the tester always rejects the module (provided only that $t > 0$) and so the user cannot experience a malfunction. Note that P_e is neither the probability {user has a malfunction} nor the probability {user has a malfunction | tester accepts module}. Rather, P_e is the probability {user has a malfunction and tester accepts module}. If the number of tests is fixed at t, and if testing with replacement is done, then the value of of W which maximized P_e is

$$W_o = \frac{N}{t+1} \qquad (5)$$

From Eq. (5) or Fig. 1, it can be seen that as $W \to 0$ then $P_e \to 0$ also. This is so because for small W there is a small chance that the user will encounter a faulty input. Similarly as $W \to N$ then also $P_e \to 0$ because there is little chance that the tester will accept the module. Of course, the last case is undesirable for reasons other than the value of P_e, i.e., the user has a small probability of receiving a released program.

C. Model with Unequal Probabilities

The model described in the preceding section is too simple to apply to most practical testing situations: some inputs are more likely to fail than others, and the probability of one input failing may not be independent of another input failing. Often a single bug may cause many inputs to fail. The tester may not choose the inputs to be tested randomly, but rather in such a way as to utilize his knowledge of the a priori failure probabilities. The user may not be free to choose more reliable inputs; in fact, he may be constrained by the problem to use less reliable inputs. The model to be described here includes three events, the last two being independent of each other but dependent on the first.

(1) A programmer constructs a module which has an error pattern α with probability $\mathbf{P}(\alpha)$. α might be a binary vector $(a_1, a_2, \ldots, a_N)$ with $a_i = 1$ meaning input i fails and $a_i = 0$ meaning input i functions correctly.

(2) A tester tries certain inputs to the module and accepts the module with probability $\mathbf{R}(\text{accept}|\alpha)$. The *tester passes no* information concerning which inputs were tested to the user.

(3) A user selects one of the inputs and the module fails with probability $\mathbf{Q}(\text{fail}|\alpha)$.

The error probability defined previously is now given by

$$P_e = \sum_{\alpha \in A} \mathbf{P}(\alpha)\ \mathbf{Q}(\text{fail}|\alpha)\ \mathbf{R}(\text{accept}|\alpha) \tag{6}$$

where A is the set of all possible error patterns.

To illustrate, if there are N input values then A consists of 2^N elements. Assume $\mathbf{P}(\alpha)$ is 0 for all A except

$$a_W = (\underbrace{1, 1, \ldots, 1}_{W},\ \underbrace{0, 0, \ldots, 0}_{N-W})$$

and let the probability of the user selecting input i be q_i. Then

$$\mathbf{Q}(\text{fail}|a_W) = \sum_{i=1}^{W} q_i$$

Assume the tester selects his inputs randomly, choosing input i with probability r_i.* Then

$$\mathbf{R}(\text{accept}|a_W) = \prod_{j=1}^{W} (1 - r_j)$$

$$P_e = \left(\sum_{i=1}^{W} q_i\right) \prod_{j=1}^{W} (1 - r_j) \tag{7}$$

If $q_i = 1/N$ and $r_j = t/N$ this reduces to

$$P_e = \frac{W}{N}\left(1 - \frac{t}{N}\right)^W \tag{8}$$

Another, more useful, form for $\mathbf{P}(\alpha)$ is obtained by assuming that the ith input malfunctions with probability p_i. Then

---

*The formula also apply if r_i is given the interpretation "input i is tested and the response is noted to be wrong by the tester." Specifically $r_i = 0$ might mean either that input i was not tested, or that it was tested and an error was not detected.

$$\mathbf{P}(a) = \prod_{i=1}^{N} p_i^{a_i}(1 - p_i)^{1 - a_i}$$

$$\mathbf{Q}(f \mid a) = \sum_{j=1}^{N} a_j q_j$$

$$\mathbf{R}(a \mid a) = \prod_{i=1}^{N} (1 - r_j)^{a_i}$$

The two products can be combined in evaluating P_e:

$$P_e = \sum_{a_1} \sum_{a_2} \sum_{a_N} \sum_{j=1}^{N} a_j q_j \prod_{i=1}^{N} F_i(a_i)$$

where

$$F_i(a_i) = p_i^{a_i}(1 - p_i)^{1 - a_i}(1 - r_i)^{a_i}$$

This reduces to

$$P_e = \prod_{j=1}^{N} (1 - p_j r_j) \sum_{i=1}^{N} q_i \frac{p_i(1 - r_i)}{1 - p_i r_i} \tag{9}$$

If the tester selects t inputs deterministically, and if the inputs are permuted so that these occur first, then

$$P_e = \prod_{j=1}^{t} (1 - p_j) \sum_{i=t+1}^{N} p_i q_i \tag{10}$$

An optimum testing strategy is obtained if the inputs are permuted so that this expression is minimized. The first factor suggests testing inputs with the largest p_i, but the second factor suggests testing inputs with the largest $p_i q_i$. Let the above expression be abbreviated as $P_e = P_m P_u$ and note that P_u is the probability of the user getting an error on an untested input. Consider the effect of adding one more input to the test. Let

$$P_e' = P_m(1 - p_k)(P_u - p_k q_k)$$

$$P_e' \approx P_e\left(1 - \frac{(q_k + P_u)p_k}{P_u}\right.$$

Thus the criterion is to select the input with the largest

$$(q_k + P_u)p_k \tag{11}$$

As can be seen, this provides a weighted compromise between selection on the basis of p_k or $q_k p_k$ alone.

D. Model with Statistical Dependence

Computer programs usually fail for a whole set of input values as a result of a single bug. It is more realistic to assume that failures in different parts of the program are statistically independent rather than failures for different input values. For example, a single oversight might cause a square root program to fail for all negative numbers. Let

$$\sigma_{ij} = \begin{cases} 1 \text{ if the } j^{th} \text{ bug causes failure for input } i \\ 0 \text{ if the } j^{th} \text{ bug doesn't affect input } i \end{cases}$$

If β_j is the probability of the jth bug, if M is the number of possible bugs, and if θ_j $(j = 1, 2, \ldots, M)$ is the pattern of actual bugs, then

$$P_e = \sum_{\theta} \mathbf{P}(\theta)\ \mathbf{Q}(f|\theta)\ \mathbf{R}(a|\theta) \tag{12}$$

This is analogous to the corresponding formula in a given above. Here,

$$\mathbf{P}(\theta) \alpha \prod_{i=1}^{M} \beta_i^{\theta_i} (1 - \beta_i)^{1-\theta_i}$$

$$\mathbf{Q}(f|\theta) = \sum_{j=1}^{N} q_j \left[1 - \prod_{k=1}^{M} (1 - \sigma_{jk}\theta_k) \right] \tag{13}$$

$$\mathbf{R}(a|\theta) = \prod_{i=1}^{M} \prod_{\ell=1}^{N} (1 - r_\ell)^{\sigma_{\ell i}\theta_i}$$

Combining and rearranging gives

$$P_e = \sum_{j=1}^{N} q_j \sum_{\theta} A_j(\theta_1, \theta_2, \ldots, \theta_M) \prod_{i=1}^{M} F_i(\theta_i)$$

where

$$F_i(\theta_i) = \beta_i^{\theta_i} (1 - \beta_i)^{1-\theta} \prod_{\ell=1}^{N} (1 - r_\ell)^{\sigma_{\ell i}\theta_i}$$

and

$$A_j(\theta_1, \ldots, \theta_M) = 1 - \prod_{k=1}^{M} (1 - \sigma_{jk}\theta_k)$$

The summations over θ_i are for only two values (0,1) and can be carried out to give

$$P_e = \prod_{i=1}^{M}\left[1 - \beta_i + \beta_i \prod_{\ell=1}^{N}(1 - r_\ell)^{\sigma_{\ell i}}\right] - \sum_{j=1}^{N} q_j \prod_{i=1}^{M}\left[1 - \beta_i + (1 - \sigma_{ji})\beta_i \prod_{\ell=1}^{N}(1 - r_\ell)^{\sigma_{\ell i}}\right] \tag{14}$$

If deterministic testing is used (r = 0, 1) over the first t inputs:

$$P_e = \prod_{i=1}^{M'}(1 - \beta_i) \sum_{j=t+1}^{N} q_j \left[1 - \prod_{i=1}^{M''}(1 - \sigma_{ji}\beta_i)\right]$$

where $\prod'$ refers to a product over terms involving a value of i such that $\sigma_{ij} = 1$ for at least one value of j in the range $j = 1, 2, \ldots, t$, and where $\prod''$ refers to the other values of i. The bugs can be permuted so that $1 \leq i \leq \tau$ implies $\sigma_{ji} = 1$ for at least one value of j from 1 to t and $\tau < i \leq M$ implies $\sigma_{ji} = 0$ for all values of j from 1 to t. Then

$$\prod_{i=1}^{M'} = \prod_{i=1}^{\tau} \quad \text{and} \quad \prod_{i=1}^{M''} = \prod_{i=\tau+1}^{M}$$

The problem is then to minimize

$$P_e = \prod_{i=1}^{\tau}(1 - \beta_i) \sum_{j=t+1}^{N} q_j \left[1 - \prod_{i=\tau+1}^{M}(1 - \sigma_{ji}\beta_i)\right] \tag{15}$$

A graphical interpretation of the above is portrayed in Fig. 2, and illustrates how different inputs excite various bugs. For example, input 2 excites bugs numbered 1, 2, 3 and 4. Bug numbered M can only be discovered through the application of inputs N-3, N-2, N-1 and N.

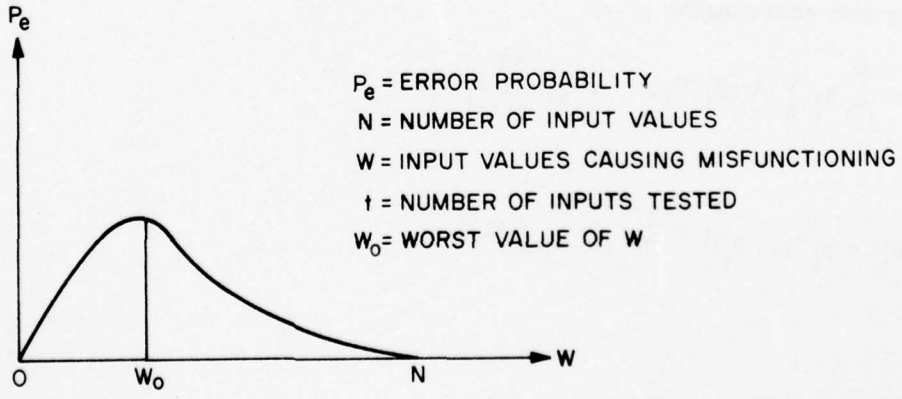


Fig. 2. Plot of P_e vs. W from the model of Equation 2.

E. Illustrative Special Cases

Some simple special cases will illustrate the above formula, and are useful in getting a rough idea of the relation between number of tests and probability of error. Suppose that each of the M bugs occurs with the same probability of error, β, and affect the same number of inputs, b. Suppose further, that the pattern of input errors is the most difficult to detect with a given number of tests, t, i.e., that the error subsets are as "disjointed as possible." If the tests are distributed most effectively over the inputs, each test will detect Mb/N bugs. This will result in M -Mbt/N undetected bugs, and τ = Mbt/N (assuming bt $\leq$ N).

$$P_e \approx (1-\beta)^{M\frac{bt}{N}}(1-\frac{t}{N})\left[1-(1-\beta)^{M(1-\frac{bt}{N})}\right] \tag{16}$$

Usually t << N and the middle term can be neglected. For the testing to be effective their number must be larger than a critical value given by

$$\left.\begin{array}{l} t > \dfrac{N}{\beta M b} \\ t > \dfrac{N}{b} \end{array}\right\} \text{ which were given a smaller value} \tag{17}$$

The latter is independent of β because the testing is exhaustive.

F. Alternative Definitions of Error

The definition of error which was used above may not be suitable for all purposes, but related probabilities can easily be calculated from the equations given. Define two events as follows:

E_u is the event that a user employs the module once and encounters a failure.

E_m is the event that a tester operates the module t times and does not encounter a failure, i.e., the tester accepts the module. The subscript m is mnemonic for "missed," since bugs are always assumed to be present with some probability, and therefore $P_r\{E_m\}$ is the probability that the tester has missed all bugs, i.e., incorrectly accepted the module.

Three different quantities which might be interpreted as the probability of user error are tabulated below for the various models which have been analyzed. The effectiveness of the testing process can be judged by comparing the first with the last two. Note that the last, $P\{E_u|E_m\}$, is always greater than the second, $P\{E_uE_m\} = P_e$. This fact might lead to a choice of the conditional probability, $P_r\{E_u|E_m\}$, in some cases where P_e is small due to tester rejection and in spite of a large probability of module bugs.

G. Conclusions and Comments

It is believed that defining what is meant by the probability of a program error, and presenting a model which permits its exact calculation (Eq. 15) will provide the nucleus around which a theory of software reliability can be built. The purpose is not merely to get a formula into which numbers can be plugged to give a probability of error -- this does nothing to reduce errors. Rather it is anticipated that by classifying different types of bugs, errors and test results, and by showing how they interact programming systems can be improved and optimum testing procedures can be found.

Table I summarizes the models discussed.

TABLE I. Summary of the statistical test models.

Probability of → / Model ↓	user failure	user failure and acceptance	user failure, given acceptance
Equal failure probabilities $p_i = \frac{W}{N}$.	$\frac{W}{N}$	$\frac{W}{N}(1 - \frac{W}{N})^t$	$\frac{W}{N}$
Unequal failure probabilities $M=N, \sigma_{ji}=\delta_{ji}$	$\sum_{i=1}^{N} q_i p_i$	$\sum_{i=t+1}^{N} q_i p_i \prod_{j=1}^{t} (1-p_j)$	$\sum_{i=t+1}^{N} q_i p_i$ (sum over untested inputs)
With statistical dependence	$\sum_{j=1}^{N} q_j \left[1 - \prod_{i=1}^{M} (1-\sigma_{ji}\beta_i) \right]$	see Eq. (15)	$\sum_{j=t+1}^{N} q_j \left[1 - \prod_{i=r+1}^{M} (1-\sigma_{ji}\beta_i) \right]$
Special case of above	$1 - (1-\beta)^M$	see Eq. (16)	$(1 - \frac{t}{N}) \left[1-(1-\beta)^{M(1-\frac{bt}{N})} \right]$

3 TYPES OF ERROR PROBABILITY

Rome Air Development Center
F-30602-74-C-0294

A.E. Laemmel

STATISTICAL THEORY OF PROGRAM TESTING AND PROVING

A. E. Laemmel

The purpose of this work is to extend results on program testing which were given in Ref. 1 and to discuss the efficiency and statistics of combined testing and proving programs. A program, call it an analyzing program, which is designed to test and/or prove another (subject) program is prone to many practical and theoretical difficulties. In fact, if one ponders the process of testing and proving the analyzing program itself, one quickly arrives at the conclusion that a perfectly general version cannot exist. Nevertheless, computer programmers are constantly being given subject programs about which they must say something, and subject programs are constantly given to users without being completely tested or proved correct. Note that we must, for practical reasons, leave open the possibility that part of the analyzing program is carried out inter actively by the human operator.

More specifically, it will be shown how better analyzing programs can be written by:

(1) Using several criteria for optimum search strategies to minimize the probability of indecisive outcomes

(2) Imitating human reasoning processes, especially in seeking proof

(3) Using previously reported statistical results to provide more realistic probabilities for (1) above.

Before treating testing and proving programs a class of "decision programs" will be discussed. These differ from most computer programs in what the answer is not "105.3" or "yes" -or- "no," but rather "yes" -or- "no" or "I don't know." The latter case of indecision is not a result of inability in programming ability, rather it is intrinsic in the problem.

A. Decision Programs

For the purpose of the present discussion, a decision program will be defined as one which examines a given object or set of objects and which outputs one of several messages. This definition is purposely too general (it includes almost every program), but it will be narrowed by example and by definition below. It does not seem desirable to exclude programs such as one for calculating $\tan(\theta)$; these can meet the definition as degenerate cases. An example which illustrates the idea of a decision program is one which examines an initial string such as 000101 and which outputs the result of applying Post's tag rules (0..S→500, 1..S→51101). Possible outputs might be:

(1) THE STRINGS VANISH AFTER n ITERATIONS

(2) THE STRINGS ULTIMATELY REPEAT WITH A PERIOD p

(3) ITERATION LIMIT t EXCEEDED

(4) MEMORY ALLOCATION m EXCEEDED.

A SNOBOL program which does this is given in Section C. Note a characteristic feature of most decision programs is outcomes (3) and (4). They say essentially that no decision was reached within the allocated bounds.

Another example of a decision program is a testing program which examines another subject program. Possible outputs here might be:

(1) SUBJECT PROGRAM FAILS ON INPUT x

(2) SUBJECT PROGRAM OPERATED CORRECTLY ON n DIFFERENT INPUTS

Note here that, except for the very simplest subject programs, to expect an output such as

(2a) SUBJECT PROGRAM WILL ALWAYS OPERATE CORRECTLY

is both theoretically and practically unreasonable. One can make outcomes (3) and (4) in the first example very unlikely by setting t and m large enough, but there is no way to calculate values of t and m ahead of time which will permit (3) and (4) to be replaced by

(3a) THE STRINGS GROW WITHOUT BOUND

We are thus faced with the fact that in many cases of interest, some of the possible outcomes of a decision program have the practical interpretation "NO DECISION COULD BE REACHED."

Proof/counter example model. A special case of a decision program occurs very frequently in various fields of mathematics and, in particular, it describes the situation in computer program testing. Abstractly, a formula of the form

(x) T(x)

is given, meaning that for every x in the universe of discourse the predicate T is true for the subject x.

Two examples are:

Example 1: (x) means "all even integers greater than 2" Goldbach's conjecture T(x) means "x is the sum of two primes."

Example 2: (x) means "all non-negative integers" program testing T(x) means "the program correctly calculate x!"

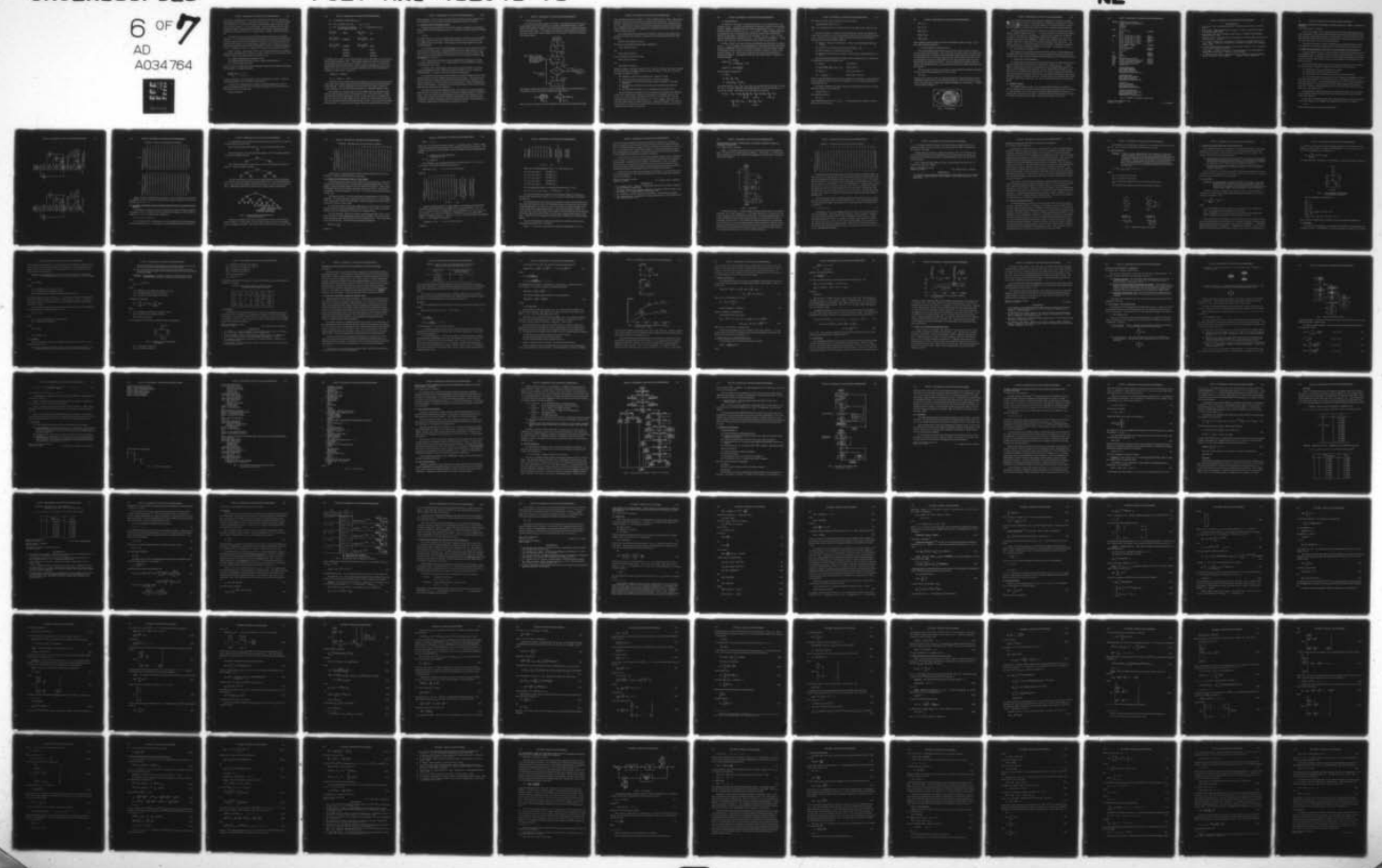
The abstract model will now be enlarged to provide a convenient formalism. We are given (x) T(x) and are to say that it is either correct or incorrect, i.e., one of the following:

$$M_1 M_2 \ldots M_r L_1 L_2 \ldots L_m \vdash (x)\ T(x) \qquad (1)$$

$$(\exists y)\ \overline{T}(y) \qquad (2)$$

AD-A034 764 POLYTECHNIC INST OF NEW YORK BROOKLYN MICROWAVE RESE--ETC F/G 20/3
PROGRESS REPORT NUMBER 41, 15 SEPTEMBER 1975 THROUGH 14 SEPTEMB--ETC(U)
NOV 76 A A OLINER F44620-74-C-0056
UNCLASSIFIED POLY-MRI-452.41-76 NL

6 OF 7
AD
A034764



The first is a symbolism for the fact that a proof has been found for (x) T(x) starting with algebraic axioms $M_1 M_2 \ldots M_n$, logical axioms $L_1 L_2 \ldots L_m$, and using accepted rules of inference. The second is simply a statement that a subject y has been found for which T(x) is false, i.e., a counter example. Note that the axioms M_i are called "algebraic" merely to distinguish them from the logical axioms; they pertain to whatever field the Theorem T is in, e.g., number theory, analysis, program translation, etc.

In mathematics one is often faced with alternately trying to prove a suspected theorem, i.e., to alternately working on the possibilities of outcomes (1) and (2) above. Of course, there is also a third possibility that neither (1) nor (2) has been established within the allocated time and resource limits. The best that can be done is to try to minimize the probability of this third "NO DECISION" outcome, to minimize the expected cost to a definite decision, or to optimize with respect to some similar parameter.

Black's strategy. The simplest version of the situation described above might be rephrased in traditional terms as follows: a single red ball is in one of two boxes, each of which contains a very large number of white balls. Let

p_i = a priori probability that red ball is in box i

m_i = probability that if red ball is in box i a single look will miss it

c_i = cost of a single look in box i

Black[1] has given a simple way to find the minimum expected cost strategy for searching for the red ball. Arrange the numbers

$$\frac{p_i m_i^n (1 - m_i)}{c_i} \quad \begin{cases} i = 1, 2 \\ n = 1, 2, 3, \ldots \end{cases}$$

in decreasing order. If the k-th number in this arrangement is one with i = 1 then the k-th look should be in box i, otherwise in the other box.

In order to get a feeling for this strategy in the present application, let box 1 represent possible proofs that a certain computer program works correctly and box 2 represent various sets of input data.

Choose parameters:

$p_1 = .9 \leftarrow$ (program probably OK) $\rightarrow p_2 = .1$

$m_1 = .999$, $c_1 = 100$ } computer search for a proof time-consuming and not likely to succeed

$m_2 = .2$, $c_2 = 5$ } one program test is easy

$\frac{p_1(1-m_1)}{c_1}$	$= .000009$	$\frac{p_2(1-m_2)}{c_2}$	$= .016$
$\frac{p_1(1-m_1)m_1}{c_1}$	$= .00000899$	$\frac{p_2(1-m_2)m_2}{c_2}$	$= .0032$
$\frac{p_1(1-m_1)m_1^2}{c_1}$	$= .00000898$	$\frac{p_2(1-m_2)m_2^2}{c_2}$	$= .00064$
	$.00000897$		$.00128$
	$.00000896$		$.0000256$
	$.00000895$		$.00000512$

According to this, looks 1, 2, 3, 4 and 5 should be in box 2, then looks 6, 7, 8 (for several hundred looks) should be in box 1. This strategy simply says to look in the box which has the highest ratio of a posteriori probability to cost at any stage. It is intuitively satisfying that the first looks were for counter examples, and (perhaps) that the sixth look was for a proof. It is not too satisfying that these looks are to be followed by a attempts at a proof, where a is given by

$$.000009\, m_1^a = .00000512$$

$$.57 = (1 - .001)^a \approx 1 - .001\, a$$

If the first series above decreased more rapidly, and the second series decreased more slowly, the strategy would go back and forth between proof and counter example more frequently. The main difficulty is the successive trials, either for proof or counter examples, are not statistically independent as is required in Black's model.

Each attempt at a proof can build on the last attempt because any partial result obtained in the previous attempt can be used in the next. This means that m_1 should decrease with n instead of remaining constant. On the other hand, if a computer program is tested with 10 random inputs, and if no failure occurs for the first 9, then the 10th trial certainly gives less information than the 1st trial. This means that m_2 should increase with n, i.e., that errors are harder to find with a single test later on in the

testing sequence. Such a result can be derived from formulas given previously,[8] since these represent a way to describe statistical dependence among test outcomes.

B. Automatic Search for Program Proofs

In the process described above it is easy to see how a computer can be programmed to look for counter examples, but how is a computer to look for a proof of a formula such as

(x) T(x)

For infinite universe, and even for large finite universes, the falsity can be established by a single value for which $T(x_i)$ is false, but no number of x_i for which $T(x_i)$ is true can establish truth for all values. At least three ways have been explored to circumvent this difficulty:

(1) In the Herbrand search method[6] single instance of a more complicated formula serves to prove T(x) for all x. The difficulty here is that this single instance is very hard to find much more so than a typical instance which causes T(x) to be false. Note that a basic cause of this difficulty is that the search is for a proof of the program and the algorithm it is based on.

(2) A related search procedure has been studied which converges to x T(x) by "successive approximations," i.e., a weaker theorem is first established, and then it is strengthened at each step.

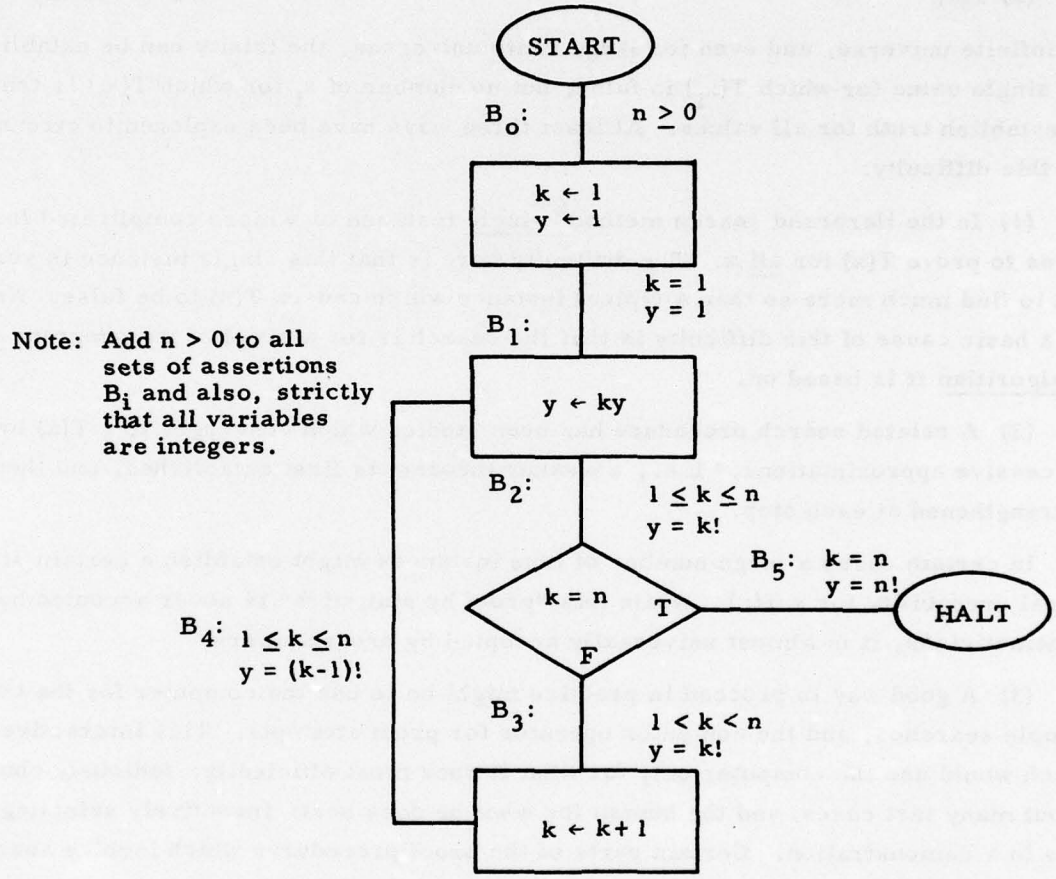
In certain cases a large number of true instances might establish a certain statistical credability for x T(x). While this "proof by statistics" is never accepted by mathematicians, it is almost universally accepted by programmers.

(3) A good way to proceed in practice might be to use the computer for the counter example searches, and the computer operator for proof attempts. This interactive approach would use the computer only for what it does most efficiently: tediously checking out many test cases; and the human for what he does best: inventively selecting the steps in a demonstration. Certain parts of the proof procedures which involve searches might also be carried out by the computer.

Four sections follow which explain (Section 1.) the inductive assertion method of program proof. This leads to a theorem to be proved, usually by (Section 2.) Herbrand search. Here the connection between Herbrand search proofs and mathematics "textbook proofs" is shown in the hope of improving the efficiency of the former (Section 3.) an alternate method for generating proofs by computer is sketched. Finally (Section 4.) a method for optimizing the required decision programs is given.

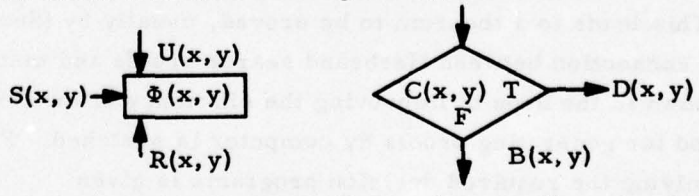
1. Inductive Assertion

This method of proving program correctness, as distinguished from merely testing the program with certain input data, is usually called Floyd's Method.[2] A brief history of the method, an example, and a discussion of its basis in mathematical induction will be found on pgs. 11-21 of Knuth's first volume on *programming*.[3] Good discussions and reviews can be found in Elspas et. al.,[4] and London.[5] Consider the following very simple example:



Note: Add $n > 0$ to all sets of assertions B_i and also, strictly that all variables are integers.

This example computes the factorial, but the method applies to any program which can be reduced to a flow chart with the following two types of boxes



Here each of S, U, R, W, B, C and D are predicates (or usually sets of predicates) which

make assertions about the variables x, y. These variables take on various values during program execution and for any given set of values each predicates S. . . D in true (T) or false (F). In particular C true or false determines the exit from the conditional branch box. The symbol $\Phi(x,y)$ represents the action of executable statements in the program on the variables, but it too can be regarded as a predicate which states that the stated action has been carried out.

It is required that the assertions at the output of a square box be implied by the assertions on one arrow or the other entering the box and operation performed within the box:

$$[S(x,y) + U(x,y)]\ \Phi(x,y) \supset R(x,y)$$

One is not to be confused by the or, this is equivalent to

$$S(x,y)\ \Phi(x,y) \supset R(x,y)$$

and

$$U(x,y)\ \Phi(x,y) \supset R(x,y)$$

for the decision box we must have both

$$W(x,y)\ C(x,y) \supset D(x,y)$$

and

$$W(x,y)\ \overline{C}(x,y) \supset B(x,y)$$

In this description x represents input data and y program variables. Note that do-nothing $\square$ boxes can always be added to avoid more than two inputs to a $\square$ box, or to avoid more than one input to a $\diamond$ box.

The algorithm is proved if the following four conditions are met:

(1) A set of assertions satisfying the above conditions is formed

(2) The assertions on the start arrow only put known conditions on the input variables

(3) The assertions on the halt arrow state that the desired answer has been obtained

(4) The algorithm terminates, i.e., it actually reaches the halt arrow.

The last condition is essential, yet as was shown by Turing there is no algorithm to test if (4) is true or false. However, in many practical problems halting can be shown, e.g., in the factorial program above k starts at 1, increases by i each time around the loop, and a halt occurs when $k > n > 0$. Also, finding a consistant set of assertions might require considerable skill for a complicated algorithm. The theorem to be proved is, in this sense, not usually known at the start, and the theorem proving method of Section 3 below can help here.

2. Herbrand Search

It is desired to determine if a suspected Theorem T can be inferred from a set of axioms $A_1, A_2, \ldots, A_m$ where all statements are quantities formulas of the predicate calculus. Godel's completeness theorem states essentially that "$A_1, A_2, \ldots, A_m \vdash T$ if and only if there is no model for $A_1, A_2, \ldots, A_m$ and $\overline{T}$." Trying to establish that there is no model for a set of statements may seem like an endless task, but Herbrand's Theorem at least offers a systematic search procedure: "$A_1, A_2, \ldots, A_m \vdash T$ if and only if a certain Boolean expression $Q_1, Q_2, \ldots, Q_r$ vanishes for some finite r. A brief sketch of how the Q_2 are to be calculated from the A_j and $\overline{T}$ will now be given.

Briefly, the idea is to replace all existantial quantifiers by functions, to express $A_1, A_2, \ldots, A_m$, $\overline{T}$ is product-of-sums Boolean form, and to generate the Q_i from these factors by substituting values from the Herbrand universe for the remaining variables. The Herbrand universe consists of constants, functions of constants, function of functions of constants, etc., the functions being those which were used to replace the existential quantifiers. The general procedure is described by Davis.[6] Consider the following example:

$$\text{axioms} \begin{cases} A_1 = (x)\,\overline{R}_{xx} \\ A_2 = (xyz)(R_{xy}R_{yz} \rightarrow R_{xz}) \end{cases}$$

$$\text{theorem } T = (xy)\,(\overline{R_{xy}R_{yz}})$$

The processed statements are

$$A_1 = \overline{R}_{ww}$$

$$A_2 = \overline{R}_{xy} + \overline{R}_{yz} + \overline{R}_{xz}$$

$$T = (E_{xy})\,R_{xy}R_{yx} = R_{ab}R_{ba}$$

The clauses are $\overline{R}_{ww}$, $\overline{R}_{xy} + \overline{R}_{yz} + R_{xz}$, R_{ab}, R_{ba} and the Herbrand universe is simply $H = \{a,b\}$. The Q_i are formed by substituting elements of H for the variables w, x, y and z in the clauses. Here there are only 20 different Q_i possible.

$$Q_1, Q_2, \ldots, Q_{20} = [R_{ab}R_{ba}\overset{\uparrow}{\overline{R}_{aa}}\,\overset{\uparrow}{\overline{R}_{bb}}(\overline{R}_{aa} + \overset{\uparrow}{\overline{R}_{aa}} + R_{aa})(\overline{R}_{aa} + \overset{\uparrow}{\overline{R}_{ab}} + R_{ab})$$

$$\text{w=a w=b} \qquad \text{x=y=z=a} \qquad \text{x=y=a z=b}$$

$$\ldots (\overset{\uparrow}{\overline{R}_{ab}} + \overline{R}_{ba} + R_{aa}) \ldots (\overline{R}_{bb} + \overset{\uparrow}{\overline{R}_{bb}} + R_{bb})]$$

$$\text{x=z=a y=b} \qquad \text{x=y=z=b}$$

This product vanishes because a subset of factors vanishes:

$$R_{ab}R_{ba}\overline{R}_{aa}(\overline{R}_{ab} + \overline{R}_{ba} + R_{aa}) = 0$$

Thus, it is established that T is a theorem which can be inferred from axioms A_1 and A_2.

The above example can be given many interpretations, among real numbers we cannot have a< b and b< a, set A cannot properly contain set B if set B properly contains A, etc. Formally, the two axioms define a transitive irreflexive ordering relation. The usual way to prove the theorem might be

(1) Assume, contrary to the Theorem T, that we can simultaneously have R_{ab} and R_{ba}

(2) A special case of axiom A_2 then says that $R_{ab}R_{ba} \rightarrow R_{aa}$

(3) But by axiom A_1 it is always false that R_{aa}

(4) Therefore (1) is false since it leads to a contradiction and the Theorem T is true.

The connection between these four steps and the vanishing of Eq. (1) can be seen by rephrasing the steps in symbolic form:

(1) $R_{ab}R_{ba}$ (Assumption)

(2) $R_{ab}R_{ba}(\overline{R}_{ab} + \overline{R}_{ba} + R_{aa}) = R_{aa}$ (Using axiom 2)

(3) $\overline{R}_{aa}R_{aa} = 0$ (Then using axiom 1)

(4) $\therefore\ R_{ab}R_{ba} = 0$ (Since other factor $\neq 0$)

Thus it is seen that Herbrand search can be regarded as a systematic, though inefficient, way to combine the axioms with various substitutions to try to establish the validity of a suspected theorem.

The above example was unusually simple in that the Herbrand universe was finite. Suppose an axiom or $\overline{T}$ is of the form

$$(x)(E_y)(z)\ R(x, y, z)$$

The reduction steps give

$$R(x, f_x, z)$$

and the Herbrand universe is $\{a, f_a, ff_a, \ldots\}$. Substituting these elements in all possible ways for x and y gives

$R(a, f_a, a)$

$R(s, f_a, f_a)$

$R(f_a, ff_a, a)$

$R(f_a, ff_a, f_a)$

$R(a, f_a, ff_a)$

Unless guided by human insight such aimless substitution results in a $Q_1 Q_2 \ldots Q_r = 0$ only after impossibly long times.

3. Alternate Method for Theorem Search

This method differs from Herbrand search in that it finds a theorem instead of merely proving a given theorem. Also, the process is certain to give some information at various steps before completion, this information being in the form of weaker forms of the find theorem. The process is patterned after what is believed to be the process gone through by a human in a similar search, and in this regard is hoped to be more efficient than the blind Herbrand search. There are difficult steps which might even require small Herbrand proofs, but this might best be done by a human interacting with the computer.

The basic idea is to have, at each step, a necessary (N) and sufficient (S) condition for some well-formed statement (F). In terms of sets

$S \subseteq F \subseteq N$

Elements not covered are in $\overline{S} \cap N$, this representing cases which meet the necessary conditions but which do not meet the sufficient conditions. The basic idea is to successively reduce this region of indecision by either enlarging S or contracting N. Consider the Venn diagram (Figure 6).

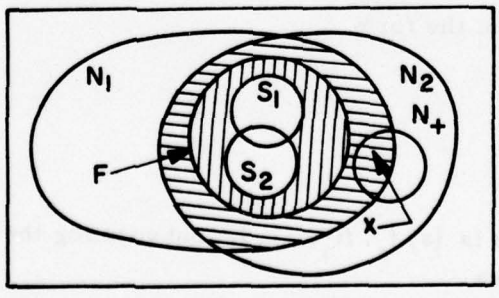


Fig. 1. Venn diagram

If N_1 and N_2 are necessary and sufficient conditions for F then the area shaded  will be empty, and if S_1 v S_2 are necessary and sufficient conditions for F then the area shaded  will be empty. If the total shaded area is empty then either combination of conditions is necessary and sufficient set, both combinations being equivalent.

The above suggests several ways to search for a theorem of the form under discussion. Suppose conditions $N_1, N_2, \ldots,$ and N_r are shown to be necessary for property F. Search for a member of (N_1 and $N_2, \ldots,$ and N_r) and F, calling such an element, if found, x. The next step is to generalize element x to set ($N_+(x)$, i.e., the set of elements satisfying the new condition N_+. Next test for a member of R and N_+. If no member exists, then enlarge the set of necessary conditions to N_1 and N_2 and ... N_r and N_{r+1} where $N_{r+1} = N_+(x)$ and repeat the whole procedure. Finally, if no further x can be found, the set of necessary conditions is sufficient. The sufficient set S_1 v S_2 ... v S_σ is expanded in an analogous way.

More details on this theorem search method will be given in a forthcoming report. Examples will be given of how to process has been applied to finite state machines and variable length codes, and how it might be applied to helping to prove computer programs.

4. Optimizing Decision Programs

In searching for theorems by the method outlined above, it is necessary in several places to quickly search for an element x satisfying n conditions $D_1, D_2, \ldots,$ and D_n. Suppose condition D_i is checked by a subroutine at cost C_i (saying in time and that condition D_i has a probability ρ_i of being true). What is the best order in which to make the tests so as to minimize their expected total cost? If all ρ_i are equal obviously the least costly should be done first, and if all costs are equal the condition with the smallest probability of success should be done first. If both ρ_i and C_i vary, then the tests should be done in order of increasing $C_i/(1 - \rho_i)$. This, and related results, were obtained by Slagle.[7]

C. SNOBOL Program

The SNOBOL program of Fig. 2 is a concrete simple example of a problem in which a complete decision cannot be expected so that time and memory limits are written in. The language is also of interest as a model for studying where the causes for observed errors were made, since there is essentially only one statement format. Notice it is about as unstructured as possible: almost every statement is a GO TO!

```
START   SYSPOT="max length was " l
        syspot="number of steps was " n
        syspot=""
        st=syspit
        l="0"
loop    l=l + "1"
        st  */l + "1"*                          /s(loop)
        dt=st
        n="1"
again   dt "0"  */"2"* *y* = y "00"             /s(a)
        dt "1"  */"2"* *y* = y "1101"           /f(nulla)
a       dt "0"  */"2"* *y* = y "00"             /s(b)
        dt "1"  */"2"* *y* = y "1101"           /f(nullb)
b       st "0"  */"2"* *y* = y "00"             /s(c)
        st "1"  */"2"* *y* = y "1101"
c       st */" 60(max length)                   /s(long)
        n=n + "2"
        n "401 (time limit)"                    /s(time)
        st */l + "1"*                           /f(equ)
        l=l + "1"
equ     st dt                                   /f(ret)
        dt st                                   /s(per)
ret     return                                  /(again)
nulla   n=n - "1"
nullb   syspot="string vanishes"                /(start)
long    syspot="string longer than max"         /(start)
time    syspot="time limit exceeded"            /(start)
per     syspot="becomes periodic                /(start)
end     syspot="end"

        100100100100100100
        BECOMES PERIODIC
        MAX LENGTH WAS 34
        NUMBER OF STEPS WAS 101

        100100100100100
        TIME LIMIT EXCEEDED
        MAX LENGTH WAS 56
        NUMBER OF STEPS WAS 401

        1010101010
        STRING VANISHES
        MAX LENGTH WAS 11
        NUMBER OF STEPS WAS 12

        10010010010010010100
        STRING LONGER THAN MAX
        MAX LENGTH WAS 59
        NUMBER OF STEPS WAS 169
```

Fig. 2. SNOBOL 3 program for Post's tag.

Rome Air Development Center
F30602-74-C-0294

A.E. Laemmel

REFERENCES

1. W. L. Black, "Discrete Sequential Search," Information and Control, 8, 159-162 (1965).

2. R. W. Floyd, "Assigning Meanings to Programs," in Math. Aspects of Computer Programming, AMS (1967).

3. D. E. Knuth, "The Art of Computer Programming," 1, Fundamental Algorithms (Addison-Wesley, 1968).

4. B. Elspas, et al., "An Assignment of Techniques for Proving Program Correctness," Comput. Survery, 4 (1972).

5. R. L. London, "Experience with Inductive Assertions for Proving Programs Correct," Symp. on Semantics of Algorithmic Languages (E. Engeler, Ed., Springer-Verlag, 1970).

6. M. Davis, "Eliminating the Irrelevant from Mechanical Proofs," Proc. 15th Symp. in Appl. Math., AMS, 20-22 (1963).

7. J. R. Slagle, "An Efficient Algorithm for Finding Minimum Cost Procedures for Making Binary Decisions," J. ACM, 11, 253-264 (1964).

8. M. Shooman and H. Ruston, "Summary of Technical Progress Software Modeling Studies," Report No. POLY EE/EP- - , Polytech. Inst. of New York

PROGRAM PATHS AND THE NUMBER OF TESTS NEEDED TO VERIFY A COMPUTER PROGRAM

G.S. Popkin and M.L. Shooman

In an earlier report[1] it was shown how matrices and binary programming may be used to determine a lower bound on the minimum number of test cases needed to verify that a program agrees with its flowchart. In this work, it is shown how flowchart contents affect the actual number of test cases needed, and a formula for the upper bound on the minimum number of cases is given and proved.

This work also includes a method of finding the actual minimum number of test cases required, using binary programming and an enumeration of all program paths.

A. Upper and Lower Bounds on the Number of Tests Needed to Verify a Program

Using the flowchart in Fig. 1, with its segment numbered, and following the methods described in Ref. 1, the matrices M and T may be constructed with k = 7 (Table I). The methods of Ref. 1 also yield 5 as the size of the maximum incomparable set, which is the lower bound on the minimum number of tests needed to pass through every segment at least once.

The actual minimum number of test cases required will depend on the contents of the flowchart boxes. Sufficient conditions for achieving the lower bound are

(1) The decisions are all independent of one another

(2) The decision variables are all read as input

(3) The program does not modify the values of the decision variables.

If, on the other hand, the decisions have a certain degree of dependence, the lower bound may not be achievable. The flowchart in Fig. 2, for example, has the same structure as that in Fig. 1 but cannot be tested with only 5 cases. Owing to the dependence of the decisions, 6 tests are needed in order to pass through every segment at least once. There are several different sets of 6 tests which will exercise every segment, but no set smaller than 6.

By changing the contents of the boxes in the flowchart in Fig. 2, it is possible to create a different flowchart (but with the same structure) that will require 7 tests to exercise all segments.

In the case of a flowchart without loops, the upper bound on the minimum number of test cases needed to pass through every segment at least once, u, is given by

$$u = d + 1$$

where d is the number of deciders in the flowchart.

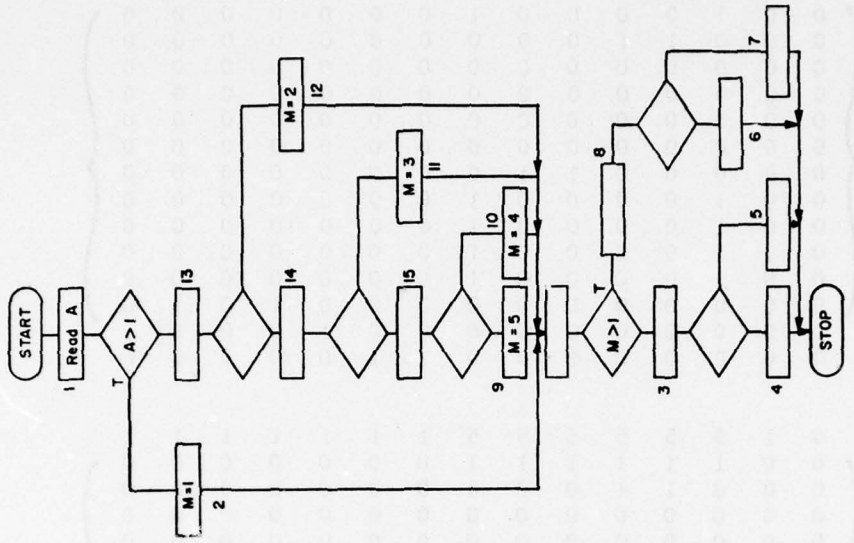


Fig. 2. The flowchart of Fig. 1 with contents.

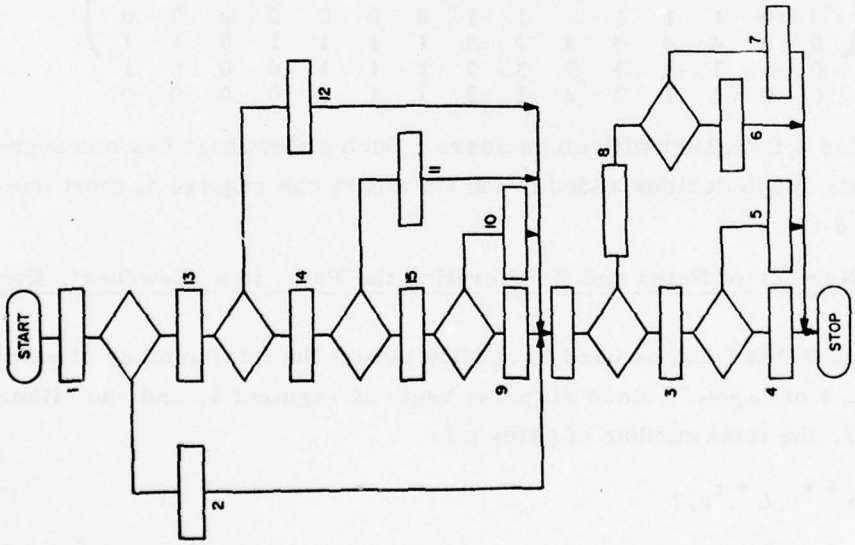


Fig. 1. A flowchart.

TABLE I. M and T for the flowchart of Figure 1.

$$M = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \\ 11 \\ 12 \\ 13 \\ 14 \\ 15 \end{array} \begin{array}{c} \begin{array}{ccccccccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 \end{array} \\ \left(\begin{array}{ccccccccccccccc} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{array}\right) \\ \begin{array}{ccccccccccccccc} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \end{array} \end{array}$$

$$T = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \\ 11 \\ 12 \\ 13 \\ 14 \\ 15 \end{array} \begin{array}{c} \begin{array}{ccccccccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 \end{array} \\ \left(\begin{array}{ccccccccccccccc} 0 & 1 & 5 & 5 & 5 & 5 & 5 & 5 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4 & 4 & 4 & 4 & 4 & 4 & 1 & 1 & 1 & 1 & 0 & 1 & 1 \end{array}\right) \\ \begin{array}{ccccccccccccccc} 0 & 0 & 3 & 3 & 3 & 3 & 3 & 3 & 1 & 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 2 & 2 & 2 & 2 & 2 & 2 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \end{array} \end{array}$$

Proof: Consider a flowchart with no deciders. Such a flowchart has one segment and requires one test. Each decider added to the flowchart can require at most one additional test, so $u = d + 1$.

B. Calculating the Number of Paths and Enumerating the Paths in a Flowchart, Using Matrices

The matrix T in Table I can be used to readily reveal the total number of paths in the flowchart of Fig. 1 of Lipow.[2] Since all paths begin at segment 1, and end either at segment 4, 5, 6 or 7, the total number of paths p is

$$p = t_{1,4} + t_{1,5} + t_{1,6} + t_{1,7}$$

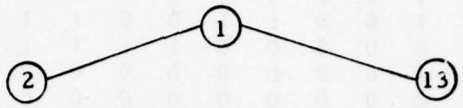
This is seen to be 20, a number which can be obtained also by the decomposition method described in Ref. 1 or by the direct calculation method described by Shooman.[3]

An enumeration of the 20 paths may be generated from the matrix M, and the enumeration takes the form of a tree.

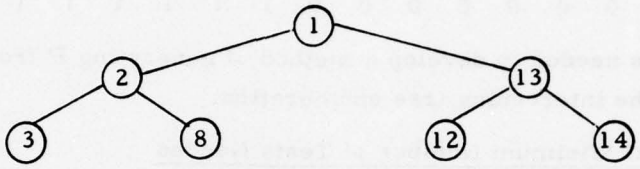
Since all 20 paths begin with segment 1, we start with the tree with node 1



From the matrix M, row 1 indicates that segments 2 and 13 can follow segment 1, so 2 and 13 are added to the tree.



Row 2 of M indicates that 3 and 8 follow 2, and row 13 indicates that 12 and 14 follow 13. These are added to the tree.



The tree is built level by level, with each bottom terminus of the tree directing to a row of M, until construction of the tree terminates due to a row of M containing all zeros. This process was applied to the flowchart of Fig. 1, and the result is shown in Fig. 3, showing the segments that comprise each of the 20 paths in the flowchart.

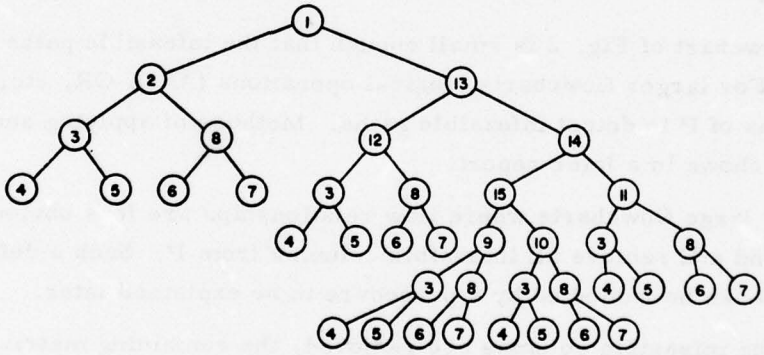


Fig. 3. A tree showing all the paths in the flowchart of Figure 2.

The paths so enumerated can now be represented by a matrix, which will prove useful in Section C. In the matrix, P, each column represents a path, and contains a 1 for each segment in the path. The matrix P from the tree in Fig. 3 is given in Table II.

TABLE II. The path matrix for the path tree in Figure 3.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0
4	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0
5	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0
6	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1
7	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0
8	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1
9	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
12	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
14	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1
15	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	0	0	0

Further work is needed to develop a method of generating P from M by matrix operations, without the intervening tree enumeration.

C. Finding the Actual Minimum Number of Tests Needed

To find the minimum number of tests needed for any given flowchart, start with the path matrix P and remove any columns which represent infeasible paths. For example, to find the minimum number of tests for the flowchart in Fig. 2, remove from P (Table II) columns which represent paths that cannot be traversed. One such path, for example, consists of segments 1, 2, 8 and 7. It cannot be traversed because of the way M is set.

The flowchart of Fig. 2 is small enough that the infeasible paths can be found by inspection. For larger flowcharts, logical operations (AND, OR, etc.) can be applied to the columns of P to detect infeasible paths. Methods of applying such logical operations will be shown in a later report.

In very large flowcharts where flow relationships are less obvious, it may not be possible to find and remove all infeasible columns from P. Such a defect is not fatal, however, and can be remedied by a procedure to be explained later.

After the infeasible columns are removed, the remaining matrix, say U, can be used to construct a binary programming problem whose solution is the number of test cases required, and the paths that those test cases should traverse. Let J be the index set of the feasible paths. Then the binary programming problem is

$$\text{minimize } z = \sum_{j \in J} x_j$$

subject to

$$UX \geq 1 \qquad x_j = 0, 1$$

where X is a vector whose transpose is $(x_j | j \varepsilon J)$. The solution yields a value for z which is the minimum number of tests needed to pass through all segments at least once. The variables x_j refer to the j feasible paths, and in the solution to the binary programming problem

$$x_j = \begin{cases} 1, & \text{if path j is one of the paths to be traversed by a test} \\ 0, & \text{otherwise} \end{cases}$$

As an example, the 10 feasible paths from P in Table II are 1, 2, 7, 8, 11, 12, 15, 16, 19 and 20. So the binary programming problem is

$$\text{minimize } z = \sum x_j \qquad j = 1, 2, 7, 8, 11, 12, 15, 16, 19, 20$$

subject to

$$\begin{array}{c|cccccccccc} & 1 & 2 & 7 & 8 & 11 & 12 & 15 & 16 & 19 & 20 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 3 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 4 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 5 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 6 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 7 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 8 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 9 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 10 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 11 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 12 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 13 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 14 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 15 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \end{array} \times \begin{pmatrix} x_1 \\ x_2 \\ x_7 \\ x_8 \\ x_{11} \\ x_{12} \\ x_{15} \\ x_{16} \\ x_{19} \\ x_{20} \end{pmatrix} \geq \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

$$x_j = 0, 1; \qquad j \varepsilon J$$

This is a standard-form set covering problem, and the matrix reductions given by Garfinkel and Nemhauser[4] may be applied to U. By their Reduction 4, $r_1 > r_2 = r_3 > r_4$, so rows 1, 2 and 3 are eliminated. Also by Reduction 4, $r_8 = r_{13} > r_{14} > r_{15} > r_9$, so rows 8, 13, 14 and 15 are deleted. By Reduction 2, $r_4 = e_1$, so row 4 and column 1 are deleted, and $x_1 = 1$. Also by 2, $r_5 = e_2$, so row 5 and column 2 are deleted, and $x_2 = 1$. The problem that remains is

$$\text{minimize } z = \sum x_j \qquad j = 7, 8, 11, 12, 15, 16, 19, 20$$

subject to

$$
\begin{array}{c} \\ 6 \\ 7 \\ 9 \\ 10 \\ 11 \\ 12 \end{array}
\begin{array}{c}
\begin{array}{cccccccc} 7 & 8 & 11 & 12 & 15 & 16 & 19 & 20 \end{array} \\
\left(\begin{array}{cccccccc}
1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\
0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\
0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0
\end{array}\right)
\end{array}
\times
\begin{array}{c}
\left(\begin{array}{c}
x_7 \\ x_8 \\ x_{11} \\ x_{12} \\ x_{15} \\ x_{16} \\ x_{19} \\ x_{20}
\end{array}\right)
\end{array}
\geq
\left(\begin{array}{c}
1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1
\end{array}\right)
$$

$$x_j = 0, 1; \qquad j \epsilon J$$

There are 14 optimal solutions, with $z' = 4$. Some of them are

$x_7 = x_{11} = x_{15} = x_{20} = 1$ all other $x_j = 0$

$x_7 = x_{11} = x_{16} = x_{19} = 1$ all other $x_j = 0$

$x_7 = x_{11} = x_{16} = x_{20} = 1$ all other $x_j = 0$

$x_7 = x_{12} = x_{15} = x_{19} = 1$ all other $x_j = 0$

$x_7 = x_{12} = x_{15} = x_{20} = 1$ all other $x_j = 0$

The corresponding solutions to the original problem have $z = 6$, and

$x_1 = x_2 = x_7 = x_{11} = x_{15} = x_{20} = 1$ all other $x_j = 0$ etc.

This says that 6 tests are required to pass through all segments at least once, and the 6 tests should traverse paths 1, 2, 7, 11, 15 and 20 as those paths are defined in the matrix U.

As mentioned earlier, the matrix U may sometimes contain some infeasible paths, if the flow relationships in P were not sufficiently obvious for the detection and removal of all infeasible paths. In such a case, one or more infeasible paths may appear in the solution of the binary programming problem. Then, when an attempt is made to construct test data to traverse the paths in the solution, the infeasible paths will be detected. At that time they can be removed from U, and the binary programming problem solved again. The process can be repeated if necessary until a solution free of infeasible paths appears.

The essential differences between the procedure described here and the binary programming problem of Ref. 1 are these:

In Ref. 1, a matrix F was used to find a maximum incomparable set. F is a

square matrix whose dimension is equal to the number of segments in a flowchart. Even in a large flowchart the number of segments might not be more than several hundred, and usually the number of segments is much smaller and manageable in a practical way. F was used to form a maximization problem, in which the solution is the maximum incomparable set of segments, and also the lower bound on the minimum number of tests needed to pass through each segment at least once.

In this report, a matrix U is used to find a minimum number of tests. In general, U is not square. The number of rows equals the number of segments in the flowchart, but the number of columns equals the number of feasible paths. Even in a modest-sized flowchart, the number of feasible paths may be very large and not easy to find. U is used to form a minimization problem, and the solution is the minimum number of feasible paths needed to pass through ("cover") every segment at least once.

Further work is needed to develop methods of deriving U in a manner that is practical for real flowcharts. Also later an attempt will be made to apply these methods to flowcharts with loops.

Rome Air Development Center
F30602-74-C-0294

G.S. Popkin and M.L. Shooman

REFERENCES

1. G.S. Popkin and M.L. Shooman, "Testing and Verification of Computer Programs," Research Report in preparation.
2. M. Lipow, "Application of Algebraic Methods to Computer Program Analysis," Report TRW-55-73-10, TRW Software Series (May 1973).
3. M.L. Shooman, "Analytic Generation of Test Data," unpublished memorandum.
4. R.S. Garfinkel and G.L. Nemhauser, "Integer Programming," 302-303 (New York: John Wiley and Sons, 1972).

APPLICATION OF LOGICAL OPERATIONS TO FINDING FEASIBLE PATHS IN A PROGRAM FLOWCHART

G.S. Popkin and M.L. Shooman

In Ref. 1, a flowchart was given in which it was desired to find all the feasible paths. The flowchart is reproduced here as Figure 1. Also in Ref. 1, a matrix P was shown, displaying all the paths, feasible and infeasible, in the flowchart. P is reproduced here as Table I.

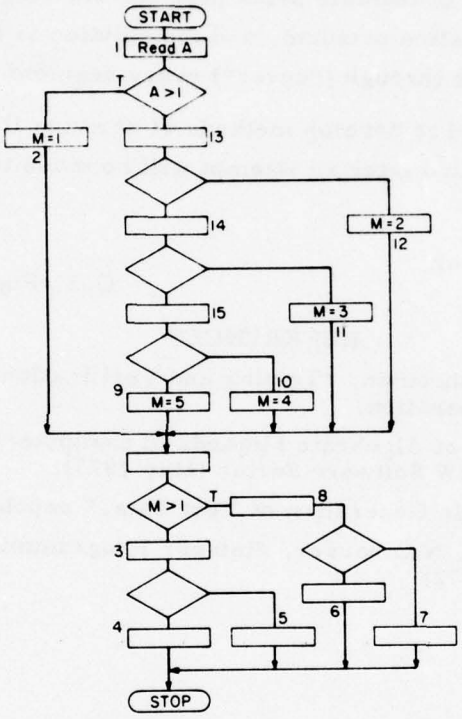


Fig. 1. A flowchart.

The flowchart in Fig. 1 is small enough so that the feasible paths can be found by inspection, as was done in Reference 1. For larger flowcharts, it is desirable to have a more methodical way of finding the feasible paths. The method proposed in Ref. 1 is to find the infeasible paths by logical operations upon P, and then to remove the corresponding columns from P. If the method of finding infeasible paths was imperfect, as it promises to be, and some infeasible paths remained undetected, then the remaining matrix would contain not only all the feasible paths but also some infeasible ones. The infeasible paths would then be found and removed by methods described in Ref. 1, as the matrix remaining from P was used in further reductions.

TABLE I. The path matrix for the flowchart in Figure 1.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1	1	0	0	1	1	0	0	1	1	1	0	1	1	0	0	1	1	0	0
4	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0
5	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0
6	0	0	1	0	0	0	1	0	0	0	1	0	0	1	1	0	0	0	1	0
7	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1
8	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
9	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1
12	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
14	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1
15	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	0	0	0	0

The approach taken in finding columns of P which represent infeasible paths, and which failed, was to establish vectors describing characteristics of feasible paths, and then performing logical operations between the vectors and columns of P. For example, from the flowchart it can be seen that any feasible path containing segment 2 must also contain segment 3, and conversely. This can be expressed as a column vector C_1 whose transpose is $C_1' = (0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)$. Also, any feasible path containing segment 12 must also contain segment 8, but not conversely. Perhaps this could be expressed as $C_2' = (0,0,0,0,0,0,0,1,0,0,0,1,0,0,0,0,0,0,0,0)$. In any case, the tester of the program could construct a number of such vectors describing relationships existing in the flowchart.

It may be possible to perform ANDing operations between the C_i and the columns of P and to find columns which represent infeasible paths. For example, the following was tried and found wanting: Any column of P, say k_j, represents an infeasible path if and only if

$$C_i \cdot k_j \neq C_i \quad \text{all } i \tag{1}$$

Using just C_1 and C_2, it so happens that Eq. (1) would correctly detect paths 3, 4, 5 and 6 as infeasible. It would incorrectly call paths 11 and 12 infeasible. Equation (1) could be made to work for paths 11 and 12 by adding to the set of the C_i a vector which says that any path containing segment 9 must also contain segment 8, namely $C_3' = (0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0)$. But for this procedure to be workable it cannot depend on the set of the C_i being complete. While it is not fatal if the procedure leaves some infeasible paths undetected, it must not under any circumstances incorrectly call a feasible path infeasible.

In an attempt to remedy the defect involved in needing a complete set of the C_i, ANDing with single vectors C_i instead of the complete set was tried. Unfortunately, this approach also failed.

Another approach would be to change the inequality in Eq. (1) to an equality and search for feasible paths instead of infeasible ones, but there still is the risk in that method that some feasible paths would be called infeasible because of the set of the C_i being incomplete.

Also, an approach using ORing was tried, with results that were even less acceptable than those from ANDing.

Rome Air Development Center
F30602-74-C-0294

G.S. Popkin and M.L. Shooman

REFERENCES

1. G.S. Popkin, "Program Paths and the Number of Tests Needed to Verify a Computer Program," Progress Report No. POLY-EE/EP-76-003, RADC-TR-76,143, 40-49 (May 1976).

PROPOSED MEASURES FOR THE EVALUATION OF SOFTWARE

S.N. Mohanty and M. Adamowicz

Comprehensive testing of software is the key route to achieve increased software reliability. If good measures for testing coverage can be defined, systematic exercise of software can lead to real improvement in the near term.[1] To achieve this testing measure the basic approach is to partition each complex module into a series of smaller, manageable internal modules, each of which is dealt with separately. After the internal modular structure of the program has been identified, appropriate aggregation of modules are made and the iteration structure for each is identified. The computation is performed as follows. The hierarchy of program sub-schema is found by automatically performing a series of reductions namely series, parallel and self-loop.[2,3] The main objective of deriving the sub-schema structure is to identify self-loop reduction points and developing a search strategy for minimizing the combinatorial growth within the backtracking process. Then the real program is analyzed by appropriate techniques so as to identify conditions which might be met within a program input-space to satisfy certain test objectives.

This report is concerned with the definition and development of two measures for characterizing the structural quality of computer programs. These measures are then used to define and measure a program's testedness (that is, how well a program is tested after a series of tests). Examples have been included to clarify and demonstrate the definitions and show how such measures can be used to study and compare different programs. It should be remembered that the testedness measures applies to the program itself and not to what the program should have been. Thus this measure might be very different from a reliability measure in the usual sense.

A. Software Evaluation Measures

A computer program consists of executable modules which are either segments or nodes. Segments consist of a contiguous set of assignment and unconditional execution statements and nodes consist of conditional or branch statements. The nature and the logical distribution of the segments and nodes of a computer program determine its particular structural qualities. In order to identify and quantify such qualities, we will introduce some new terminology. The term accessibility is introduced for quantifying the concept of reaching to or accessing an executable module during program testing. It is also assumed that before an executable module is accessed we have successfully executed the preceding modules.[4] The term testability has been introduced to quantify the ease of executing a program module. Primary factors determining the testability of a module are its accessibility and complexity measures. The complexity measure determines how much resources are required in executing the module.

The testedness concept has been introduced so as to evaluate the software testing process. It is a function of testability and frequency of execution.

B. Logical Structure

Definition: A logical structure (abbreviated to LS) is defined as a collection of at least two nodes, one of which is a start node through which the control enters the LS from the outside environment and the other being a stop node through which the control leaves the LS to the outside environment. The two nodes are connected by a directed line segment. The nodes are arranged in hierarchial levels, start node being in level 1 and the levels of the succeeding nodes being determined by the largest count of the non-repeating line segments to the node from the start node.

$$LS = \{(N_{ij}, l_{ijkl}, N_{kl}) \mid i, j, k \geq 1,\ k > 1\}$$

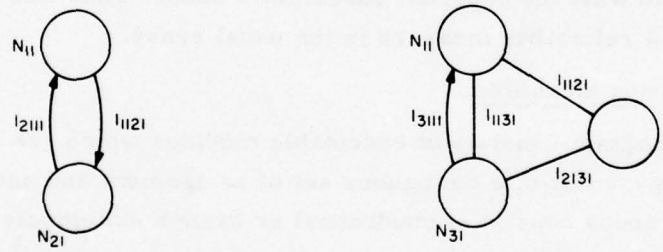
where

N_{ij} = j-th node in the i-th level,

N_{kl} = l-th node in the k-th level,

l_{ijkl} = directed line segment from the j-th node in the k-th level.

Two examples of the logical structures are portrayed in Figure 1.



START NODE - N_{11}
STOP NODE - N_{21}

$LS = [(N_{11}, l_{1121}, N_{21}), (N_{21}, l_{2111}, N_{11})]$

START NODE - N_{11}
STOP NODE - N_{31}

$LS = [(N_{11}, l_{1121}, N_{21}), (N_{11}, l_{1131}, N_{31}), (N_{21}, l_{2131}, N_{31}), (N_{31}, l_{3111}, N_{11})]$

Fig. 1. Examples of logical structures.

C. Computer Program Equivalence to Logical Structure

Any computer program can be thought of as a collection of modules, each module containing one or more executable statements, connected together directly through transfer statements (goto, call) or indirectly when modules occur in sequence. The computer program has the following properties:

(1) Each program has a start node and a stop node. Start node is the statement which signals the beginning of execution (i.e., the control enters the program environment) and the stop node is the statement which signals the completion of execution (i.e., the control leaves the program environment).

(2) Each program module can be arranged in a hierarchial level and connected to other modules through directed line segments (i.e., direct or indirect transfer statements).

If we can replace each module and the associated transfer statements by an equivalent node and a line segment then the resulting structure is a LS. Thus any computer program can be represented by an equivalent LS.

D. Accessibility

Definition: The accessibility of an executable module is defined as a sum of the products of the accessibilities of the preceding modules, times the probability of traversing the path from the preceding module to the module under consideration and the probability of successful execution of the preceding module. The accessibility of the start module is 1.

Consider a program consisting of many modules, that is, segments which are collections of sequentially executable statements and nodes which are either labels or conditional statements. Then

$$A_{kl} = \sum_{\ell_{ijkl}=1}^{i} \sum^{j} * Q_{ijkl} * P_{ij}$$

where

A_{kl} = Accessibility of the l-th module in the k-th level, under consideration

A_{ij} = Accessibility of the j-th module in the i-th level

Q_{ijkl} = Probability of traversing the path from module (ij) to the module (kl)

P_{ij} = Probability of successful execution of module j in level i.

If we assume that the probability of successful execution of any segment or node is inversely proportional to the number of basic executable statements contained in the node or segment where the basic statement is of the form VARIABLE :: = VARIABLE OPERATOR VARIABLE or VARIABLE :: = OPERATOR VARIABLE. Then the probability of successful execution of the node or segment, P_{ij}, is inversely proportional to

C_{ij}, where C_{ij} is the number of basic executable statements in the segment or node. Then, $P_{ij} = k/C_{ij}$, where k = constant of proportionality. As a special case, if k = 1 then the accessibility has the following form

$$A_{kl} = \sum_{\ell_{ijkl}=1}^{i} \sum^{j} (A_{ij}/C_{ij}) * Q_{ijkl}$$

As an example of the calculation of accessibilities, consider the flowchart of Figure 2.

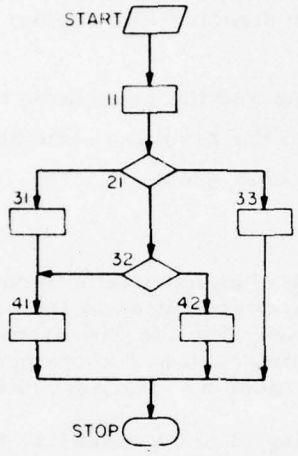


Fig. 2. A flowchart for the illustration of accessibility calculations.

The accessibilities of the modules are:

$A_{11} = 1$
$A_{21} = 1$
$A_{31} = 1/3$
$A_{32} = 1/3$
$A_{33} = 1/3$
$A_{41} = A_{31} + A_{32}/2 = 1/3 + 1/6 = 1/2$
$A_{42} = 1/6$
$A_{stop} = A_{41} + A_{42} + A_{33} = 1/2 + 1/6 + 1/3 = 1$

(Note: We have assumed that each module has one basic executable statement.)

E. Testability

Once we have computed the accessibilities of the modules, we then compute the module testabilities. Conceptually, it is the ease with which a module can be accessed

and executed in any test sequence. It is the function of the module accessibility and the resources required to execute the module. The resources can be proportional to the number of basic executable statements in the module. It can also be time and money required to execute the module.

Definition: The testability of an executable module is defined as a ratio of the module accessibility and the resources required to execute the module.

Then,

$$T_{ij} = A_{ij}/R_{ij}$$

where,

T_{ij} = Testability of the module j in level i
A_{ij} = Accessibility of the module j in level i
R_{ij} = Resources required for executing the module j in level i.

In the above expression $R_{ij} \to 0$, then $T_{ij} \to \infty$, meaning we can execute an infinite number of times without incurring any expense. On the other hand, if $R_{ij} \to \infty$, then $T_{ij} \to 0$ meaning we cannot execute the module at all because of the enormous cost involved in executing the module.

Let the resources required to execute any program module be directly proportional to the complexity of the program module. Then,

$$R_{ij} = L_{ij} * C_{ij}$$

where,

L_{ij} = constant of proportionality for module (ij)
C_{ij} = complexity of module (ij)

Then,

$$R_{ij} = A_{ij}/R_{ij}$$

or

$$T_{ij} = A_{ij}/(L_{ij} * C_{ij})$$

F. Testedness

Before giving the measures for evaluating the software in terms of the accessibilities we observe that:

(1) One has to conduct an optimum number of tests for testing the program. Increasing the number of tests beyond this optimal level should not increase

the program testedness appreciably but decreasing the number of tests below this level should decrease the program testedness appreciably.

(2) Given a fixed number of tests to be performed the net contribution towards the overall program testedness by the segments with smaller testability should be relatively large.

Definition: The testedness of a module is defined as a function of the module testability and its frequency of execution during program testing.

Then,

$$W_{ij} = 1 - e^{-(f_{ij} * T_{ij})}$$

where,

f_{ij} = Frequency of execution of the module j in level i

T_{ij} = Testability of the module j in the level i

W_{ij} = Testedness of the module j in the level i.

Testedness of a program.

$$\overline{W} = \sum_{\ell_{ijkl}=1}^{i} \sum^{j} (P_{ij} * W_{ij}) / \sum_{\ell_{ijkl}=1}^{i} \sum^{j} (P_{ij})$$

where

P P_{ij} = Probability of execution of module j in level i

W_{ij} = Testedness of the module j in level i

$\overline{W}$ = Program testedness.

As an example of these concepts consider the flowchart of Figure 3.

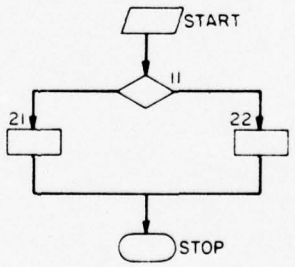


Fig. 3. Flowchart for the illustration of testedness.

T_{21} = Testability of module 21

T_{22} = Testability of module 22

f_{21} = Frequency of execution of module 21
f_{22} = Frequency of execution of module 22
W_{21} = Testedness of module 21
W_{22} = Testedness of module 22
$\overline{W}$ = Program testedness
$P_{21} = P_{22} = 0.5$

The following table shows the program testedness for different values of testability and frequency of execution.

TABLE I. Program testedness for different values of testability and frequency execution.

T_{21}	T_{22}	f_{21}	f_{22}	W_{21}	W_{22}	$\overline{W}$
0.5	0.5	1	10	0.393	0.993	0.693
0.5	0.5	10	10	0.993	0.993	0.993
0.5	0.5	10	100	0.993	1.0	0.9965
0.001	0.12	1	29	0.001	0.97	0.485
0.001	0.005	1200	1200	0.699	0.997	0.848

G. Conclusion

In this report we have introduced some new concepts and proposed measures for evaluating software. Usefulness of testability and testedness has been demonstrated by examples. The examples included are rather small but the concepts can be extended to a program of any size. These concepts can be used to automatically test a computer program and evaluate its testedness.

Rome Air Development Center
F30602-74-C-0294

S.N. Mohanty and M. Adamowicz

REFERENCES

1. E.F. Miller, Jr. and R.A. Melton, "Automated Generation of Testcase Datasets," Proc. Intl. Conf. on Reliable Software, 51-58 (April 1975).
2. E. Ashcroft and Z. Manna, "The Translation of GOTO Program GO WHILE Program," in Information Processing (North Holland Publishing Co., 1972).
3. J.E. Sulivan, "Measuring the Complexity of Computer Software," MITRE Corp., MTR-2648 (June 1973).
4. S.N. Mohanty, "Automatic Program Testing," Ph.D. Dissertation, Polytech. Inst. of New York (June 1976).

SEEDING/TAGGING ESTIMATES OF THE NUMBER OF SOFTWARE ERRORS

B. Rudner

Previous reports have analyzed estimates supported by three different models of the debugging process. The first model assumes that all errors have equal probability of discovery. Under this assumption, the maximum likelihood estimate of N, the total number of errors in a program, is $N_o = st/c$, where t errors are either deliberately inserted in a program (seeded) or found by debugging (tagged); s errors are found by a debugger unaware of the contents of the first set; and c is the number of errors appearing in both sets. Chapman[1] shows that N_o is biased and gives simple approximate forms for the bias and mean-squared error* which are stated to be valid for the ratio $st/N \geq 10$. An analysis of a modified maximum-likelihood estimate $N_1 = \frac{(s+1)(t+1)}{(N+1)} - 1$ was given in the same paper. N_1 has the advantage of being practically unbiased.

A. Comparison of Statistics Computed in Different Ways

Other estimates have also been considered. Among them are estimates involving more than one value of c, found from several tests. It was necessary to use methods other than those used by Chapman to find the statistics of two of these multi-trial estimates, and a method based on Taylor's series was applied to find formulas for mean and mean-squared error. A by-product of this work was that by reducing the number of trials to 1 in the appropriate formulas we have alternative formulas for the statistics of N_o. It was pointed out in Ref. 2 that the results obtained in this way were considerably lower than those found previously. A second inconsistency was noted when comparison was made with some specific cases in a tabulation containing means and variances computed directly from probabilities.[3]

As a result, Chapman's simplified formulas were checked. It was found that certain approximations used in deriving the formulas introduced considerable error unless N, s and t were actually very large numbers; the condition st/N large was not sufficient. Table I lists three cases, for all of which st/N = 13.33, giving percent error in the original approximation formula for $E(N_o)$. Clearly the error decreases as the magnitudes increase. The mean-squared error formula shows an even larger deviation.

The bias resulting from the approximate formula was, for small N, not much larger than the error, indicating that both bias and dispersion might actually be considerably lower than appeared. Consequently, it was necessary to derive second-order approximations for bias and mean-squared error, more accurate than Chapman's

---

*Because of the bias, the mean-squared error $V(\hat{N}) = E[(\hat{N} - N)^2]$ rather than the variance was taken as a measure of dispersion.

TABLE I. Percent error in first approximation formula for $E(N_o)$ for several examples with $st/N = 13.33$.

(st/N = 13.33) N, s, t	% error in approximate formula for $E(N_o)$
30, 20, 20	7.9
270, 60, 60	3.5
3000, 200, 200	1.1

first-order approximations but still simpler than the exact expressions, in order to permit quick calculation, to provide insight into the manner in which bias and dispersion change with changing parameter values, and to facilitate comparison with other estimates.

On the basis of the new approximations, additional information was obtained on the manner in which bias and mean-squared error change with the parameters, information which is useful in designing an actual estimation experiment.

B. Bias

The new approximation for the expected value of N_o derived from Chapman's exact result is

$$E(N_o) \simeq st[a_1 + a_2 + 2a_3 + 6a_4 + \ldots + (m-1)!a_m] \tag{1}$$

where

$$a_1 \triangleq \frac{n+1}{(s+1)(t+1)}$$

$$a_i \triangleq a_{i-1}\frac{n+i}{(s+i)(t+i)}, \quad i = 2,3,\ldots$$

The requirements for accuracy are the following:

1. A sufficient number of terms must be included in the finite sum to leave the remainder term of the infinite sum from whhci it is drawn insignificant. Four or five have been found adequate.

2. The probability that $c = 0$ must be very small. By referring to the examples of hypergeometric distribution in Ref. 4 one sees that this occurs when the peak is far from 0, i.e., when the mean of the distribution, st/N, is large. In fact, common sense tells us that large samples (i.e., s and t large, and therefore st/N large) are almost certain to have elements in common: $P(0) \simeq 0$. $st/N \geq 3$ seems to be sufficient for accuracy unless N is very large (in which case the variance of the distribution and therefore P_o are large).

An alternative form of Eq. (1), derived by simple manipulations is:

$$E(N_o) \simeq N[k_1 + k_2 (\frac{N}{st}) + 2k_3 (\frac{N}{st})^2 + \ldots + (m-1)!\, k_m (\frac{N}{st})^{m-1}] \qquad (1a)$$

where

$$k_1 = \frac{1+1/N}{(1+1/s)(1+1/t)}$$

$$k_i = k_{i-1} \frac{1+i/N}{(1+i/s)(1+i/t)} \quad , \quad i = 2,3,\ldots$$

The quantities k_i are close to 1 and approach 1 in the limit as s,t and N increase. If we set all $k_i = 1$, we arrive at Chapman's approximate formula

$$E(N_o) \simeq N[1 + (\frac{N}{st}) + 2(\frac{N}{st})^2 + \ldots]$$

The formula resulting from the Taylor's series expansion is

$$E(N_o) \simeq N[1 + q(\frac{N}{st}) + 3q^2(\frac{N}{st})^2] \qquad (2)$$

where

$$q = (N-s)(N-t)/N^2$$

This is subject to the same caveat as Eq. (1): truncation effect and $P_o \neq 0$ are possible sources of error. Both tend to show up for small st/N; and for large values of N, s, t, i.e., values for which min(s,t) >> st/N.

The bias b, of an estimate $\hat{N}$ is defined by $E(\hat{N}) = N + b$. The quantity of greatest interest is the ratio b/N (or percent bias = b/N x 100%) since to estimate N = 100 as N = 120 is clearly a grosser error than to estimate N = 1000 as 1020.

The percent bias of N_o varies in three different ways: with size of tagged and sampled sets relative to total number of errors, quantified by the ratio st/N; with the total number of errors N; and with size of sampled set relative to size of tagged set, s/t. The nature of each variation, with the other two sources held constant, is as follows:

(1) b/N decreases as st/N increases, for N and s/t constant
(2) b/N increases with N, for st/N and s/t fixed
(3) For st/N and N fixed, b/N is greatest with s/t = 1.

Figure 1 illustrates these variations, and a fourth discussed below.

The first property states the unsurprising fact that, given a particular program, large samples produce accurate estimates. The third property says that if, in addition,

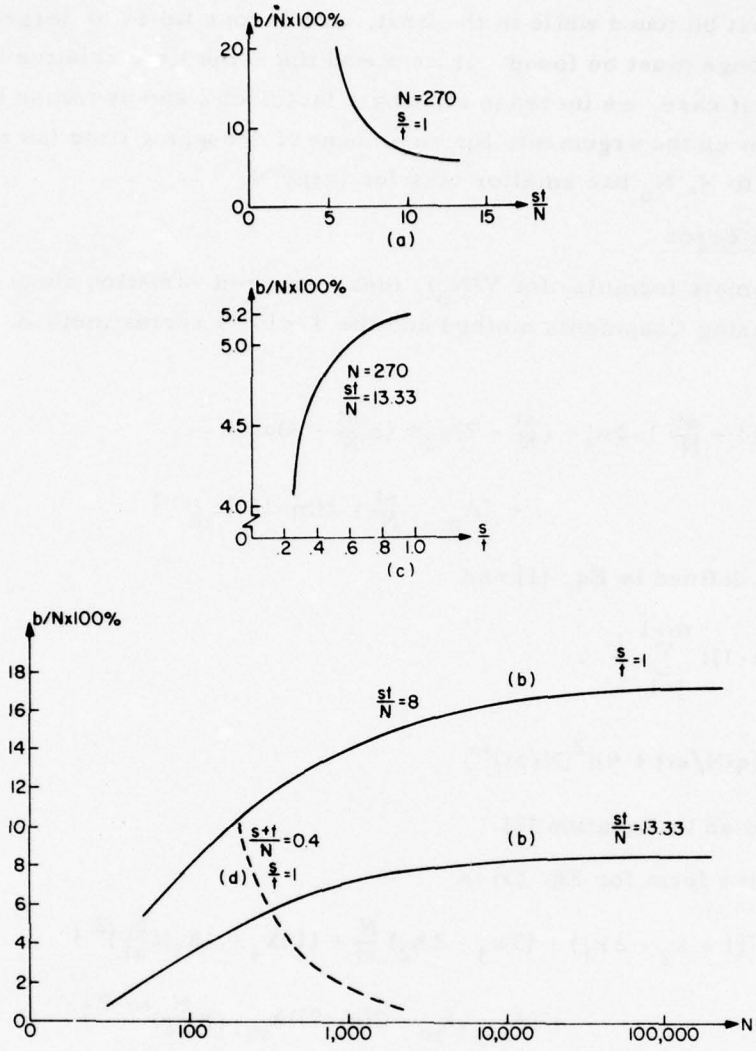


Fig. 1. Variation with several parameter relations of percent bias of N_o.

we make s and t unequal, we increase the accuracy of N_o still more. However, in both cases, the increased accuracy is paid for in time: larger sets of errors take longer to find, and s+t increases as s/t departs from 1.

The second property says that under the same conditions of st/N and s/t we get better results for programs with fewer errors, e.g., by estimating N after some debugging has been done. However, as N increases, keeping st/N constant requires relatively smaller samples. For N = 1000, for example, s = t = 100 gives st/N = 10, while for N = 250, we get the same value of st/N with s = t = 50. That is, in the second

case, 100 bugs must be found while in the first, with N four times as large, only 200, or twice as many bugs must be found. If we spend the same time relative to N and find 400 bugs in the first case, we increase st/N by a factor of 2 and decrease bias considerably. To sum up the argument, for an amount of debugging time (as measured by s+t) proportional to N, N_o has smaller bias for large N.

C. Mean-Squared Error

New approximate formulas for $V(N_o)$, mean-squared variation about the true value N, were derived using Chapman's method and the Taylor's series method. They are respectively,

$$V(N_o) \simeq N^2 \{1 + \frac{st}{N} [-2a_1 + (\frac{st}{N} - 2)a_2 + (3\frac{st}{N} - 4)a_3 + \ldots$$

$$\ldots + \{A_{m-1}\frac{st}{N} - 2(m-1)!\}a_m]\} \qquad (3)$$

where the a's are defined in Eq. (1) and

$$A_{m-1} \triangleq (m-1)! \sum_{j=1}^{m-1} \frac{1}{j} \quad .$$

$$V(N_o) \simeq N^2 [q(N/st) + 9q^2(N/st)^2] \qquad (4)$$

where q is defined as in Equation (2).

An alternative form for Eq. (3) is

$$V(N_o) \quad N^2 [(1 + k_2 - 2k_1) + (3k_3 - 2k_2)\frac{N}{st} + (11k_4 - 4k_3)(\frac{N}{st})^2 + \ldots$$

$$+ (A_{m-1}k_m - 2(m-2)!k_{m-1})(\frac{N}{st})^{m-2}] \qquad (3a)$$

where the k's are defined as in Equation (1a).

The formulas hold under the same conditions as the mean formulas: P 0, and enough terms in the series. Furthermore the same generalizations can be made with respect to the variation of $V(N_o)$ with N, s, t.

D. Modified Maximum Likelihood Estimate N_1

An intermediate result in Chapman's derivation of $E(N_o)$,

$$E(\frac{1}{c+1}) = \frac{N+1}{(s+1)(t+1)}(1-K) \ ,$$

where

$$K = \begin{cases} \frac{n-s-t}{N+1} & \text{for } s+t \leq N \\ 0 & \text{otherwise} \end{cases}$$

suggested the modified estimate

$$N_1 = \frac{(s+1)(t+1)}{c+1} - 1$$

as a means of reducing the bias to practically zero assuming $P_o \simeq 0$. For

$$E(N_1) = (s+1)(t+1)\, E(\frac{1}{c+1}) - 1 = (N+1)(1-K) - 1$$

$$E(N_1) = N - K(N+1) \quad \text{where} \quad K \simeq 0 \quad \text{if} \quad P_o \simeq 0.$$

Therefore

$$E(N_1) \simeq N$$

The bias of N_1 is negative but very small even for st/N small. For example consider the case N = 6, s = 2, t = 3, st/N = 1. $E(N_1)$, computed exactly, is 5.8, and b/N is 3.3% whereas $E(N_o)$ is 6.6 with b/N = 10%. However, for N_1 as for N_o, b/N increases if st/N is held fixed but N increased. If N = 20, s = 4, t = 5, st/N is still 1 but $E(N_1)$ is now 16.9 and b/N = 15.5%.

An additional advantage of N_1 is the fact that its variation about N is somewhat lower than $V(N_o)$ for N greater than 50. Below 50, $V(N_o)$ is smaller. The second-order approximation for $V(N_1)$, under the same approximation rules as $E(N_o)$ and $V(N_o)$ is

$$V(N_1) \simeq (s+1)^2 (t+1)^2 [a_2 + a_3 + 2a_4 + 6a_5 + \ldots + (m-2)!\, a_m] - (N+1)^2 \qquad (5)$$

$$\simeq (s+1)^2 (t+1)^2 (\frac{N}{st})^2 [(k_2 - k_1^2) + k_3(\frac{N}{st}) + 2k_4(\frac{N}{st})^2 + \ldots + (m-2)!\, k_m (\frac{N}{st})^{m-2}] \qquad (5a)$$

Let $\sigma_e = \sqrt{V}$. The variation of $\sigma_e(N_1)/N$ with relations among the parameters as described in detail in Section B is plotted in Figure 2.

E. Useful Range

It is obviously possible to make accurate and precise estimates with large enough samples; the limiting case of s = t = N produces a perfect estimate. Whether a good estimate can be made with considerably smaller samples is the issue. N_1 has almost no bias so the major problem resides in the variance (which, for zero bias, equals the mean-squared error). As Eq. (5) and Fig. 2(a) show, the variance is low for the ratio

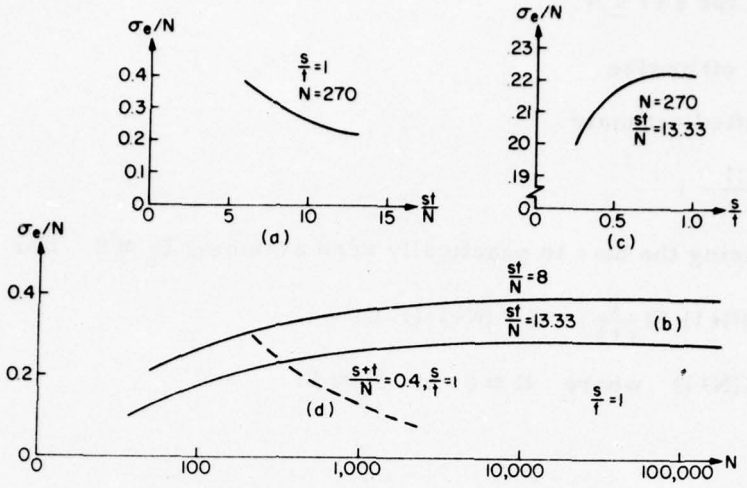


Fig. 2. Variation with several parameter relations of σ_e/N for N_1.

st/N large enough. But large ratios can be attained with relatively small samples only for N large, e.g., for N = 3000, st/N = 13.33 can be realized with s = t = 200, or one-fifteenth of N; but for N = 30, st/N = 13.33 requires s = t = 20, two-thirds of N. Fortunately, Fig. 2(b) shows that smaller values of st/N are required to give a specified value of σ_e/N at the 30-error level than at 3000. The s+t/N = 1.0 curve in Fig. 2(d) shows the minimum value of σ_e/N which can be attained if we limit s and t to half of N. If we are willing to accept larger samples, we can. of course, do better for the smaller values. Larger samples mean more time. For the same time relative to N, estimates of larger programs will have lower σ_e/N (Fig. 2(d)). Curves such as those of Fig. 2 can be exploited to design an estimation test with knowledge of the trade-off between time and precision.

F. Design of a Seeding/Tagging Reliability Test

The procedure is very simple. Our objective is to pick values for s and t which will be likely to produce an estimate of the quality we want. We begin with a ballpark estimate of the number of errors in the program, based on whatever information we have -- length of program, amount of previous debugging, experience with other programs of the same type, expertise of programmers. Suppose we decide there are probably about 150 errors. In that event N_1 is the preferred estimate since it is practically unbiased and has lower variance than N_o in that range. Had the estimated N been below 50 we should have had to check the bias and dispersion of N_o and then choose between N_o and N_1.

We will be content with $\sigma_e = 30$. Then $\sigma_e/N = .2$ and from Fig. 2(b) we find that the intersection of 150 and .2 is on the curve for $st/N = 13.33$. (If Fig. 2(a) contained a curve for N = 150, we could have found the same information there.) Then $st = 13.33 \times 150 = 2000$. We can let each debugger find about 45 errors, or let one find 50 and the other 40. Figure 2(c) shows qualitatively that the results will be about the same. We can also let s and t be, say 20 and 100 and expect a somewhat smaller σ_e but we will have to wait considerably longer for the results.

The cost beyond that for the debugging which would have to be done anyway would be identical for all choices since the additional cost is only for the common bugs and the expected number of those is $st/N = 13.33$.

The situation would be a little different if the program were not to be completely debugged. The test could, for example, be a means of comparing different programming techniques. In that case, it would not only take longer but would also be more expensive to find 120 bugs than to find 90.

Rome Air Development Center
F30602-74-C-0294

B. Rudner

REFERENCES

1. D.G. Chapman, "Some Properties of the Hypergeometric Distribution with Applications to Zoological Sample Censuses," Univ. of California Publications in Statistics, 1, 131-160 (1951).

2. M. Shooman and H. Ruston, "Summary of Technical Progress Software Modeling Studies," Report No. POLY EE/EP-75-004, SMART 100, Polytech. Inst. of New Yory, 19 (October 1, 1974-June 30, 1975).

3. M. Lipow, "Estimation of Software Package Residual Errors," TRW Systems Group, TRW-SS-72-09 (November 1972).

4. M. Shooman and H. Ruston, "Summary of Technical Progress Software Modeling Studies," Report No. POLY EE/EP-74-019, EER 114, Polytech. Inst. of New York, 15 (July 1-September 30, 1974).

MODULAR PROGRAMMING TECHNIQUES

E. Lipshitz, M. L. Shooman and H. Ruston

The objective of this effort is to find ways to write more reliable programs. To this end we raise the questions in the following three areas:

(1) Computer Language: Can we develop a high level language that is less susceptible to bugs? Can we prove that any one of the existing languages is better in the sense of having less bugs?

(2) Structural Programming and Complexity of Programs: Can we derive a sequence of steps in writing programs that will result in more reliable programs? Is there a correlation between the structure and content of a program and the number of bugs in it? If yes, which areas or statements or programming techniques are more bug-manifested? Can we eliminate or replace them?

(3) Automatic Programming: Will the use of pre-written and tested modules of code reduce the number of bugs? How can they be best incorporated into the program?

This project will investigate both structural and automatic programming with the emphasis on the latter.

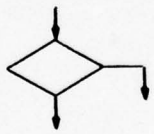
A. The Program: Auto-Programming

The program "Auto-Programming" is divided into two parts -- "Flow" and "Auto," both of which are interactive on-line programs that, by communicating with the user, generate his program. Currently, they are written and generate programs in Fortran.

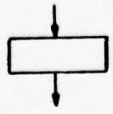
1. The Program Flow

"Flow" receives the information about the flowchart of a program from the user. "Flow" recognizes only four different types of blocks which are sufficient to generate any flowchart. They are:

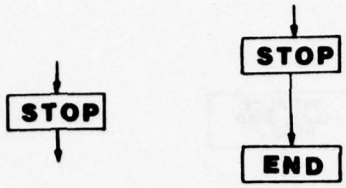
(1) Control Block. This is a conditional decision block, similar to the statement IF ( ) GO TO ___. The user will write the code for this block.



(2) Functional Block. This block will perform a task that is available in the computer library. The program "Auto" will generate the correct code for this block.



(3) Stop Block. This block indicates the end of the path; and is coded by the STOP statement.



(4) User's Code Block. The user inserts the desired code into this block. This block is used whenever the library cannot perform the needed task.

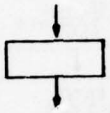


Figure 1 illustrates how the above blocks can be used to construct the flow chart of a given program. The program first builds a table "Clas" with the names of the students in a class, and then enters their grades into the table "Grad."

The flow chart generated using only the above blocks is always a binary tree in structure, while the transfer of control during the execution might behave like a graph.

Upon completing the flow chart, control passes to the program "Auto," to be discussed next.

2. The Program: Auto

The user will specify to "Auto" what he would like to do, and "Auto" will advise him which methods are available for the solution, as well as their characteristics. The user then will choose the method he prefers, and "Auto" will generate the needed code. "Auto" is divided into four parts:

(1) Dictionary, which advises the user of tasks available in the computer library.

(2) Communicant, which supplies and receives from the user all the information needed to generate the code. In addition, it will sort the labels generated by the user.

(3) Code generator, which uses information obtained with the help of the communicant, and generates the final code. It will resolve any conflicts in the labels generated by the user, and the duplicates generated by "Auto."

(4) Library, which contains a collection of code modules needed for the different tasks, e.g., input/output, solution of mathematical problems, inventory and banking programs, etc.

"Auto" was revised to allow dynamic label generation. The communicant scans the user's statements looking for labels. The labels found are stored in a table "Lab"

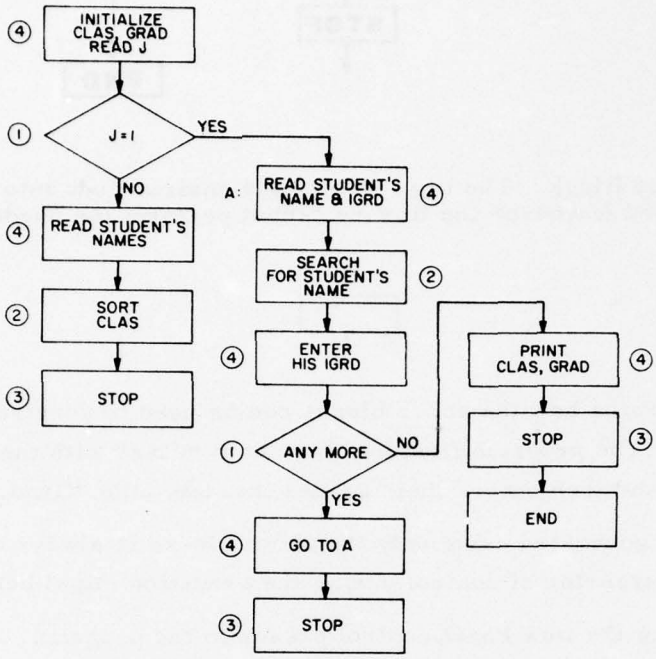


Fig. 1. Record of student's grades.

in descending order. Before the code generator writes a label it searches "Lab" to see whether this label is already in use.

The library is being expanded. Code has been written and tested for the following mathematical tasks:

$$L_n(x) = 2 \sum_{n=1,3,5}^{m} \frac{1}{n} \left(\frac{x-1}{x+1}\right)^n \qquad .5 \leq x \leq 2.0 \qquad (1)$$

$$\ell^x = \sum_{n=1}^{m} \frac{x^n}{n!} \qquad -2.0 \leq x \leq 2.0 \qquad (2)$$

$$\sin(x) = \sum_{n=0}^{m} \frac{(-1)^n x^{2n+1}}{(2n+1)!} \qquad -2\pi \leq x \leq 2\pi \qquad (3)$$

$$\cos(x) = \sum_{n=0}^{m} \frac{(-1)^n x^{2n}}{(2n)!} \qquad -2\pi \leq x \leq 2\pi \qquad (4)$$

$$\text{arc sin}(x) = x + \sum_{n=0}^{m} \frac{2 \times 4 \times 6 \ldots 2n}{1 \times 3 \times 5 \quad (2n+1)} x^{2n+1} \qquad 0 \leq x \leq .5 \qquad (5)$$

m is such that $a_m \leq 10^{-6}$ and $a_{m-1} > 10^{-6}$.

The following is an example of code generating for a program. The program is the same as in Figure 1.

Figure 2 shows the interaction needed between "Flow" and the user to generate his flow chart, as well as the flow chart generated.

Figure 3 shows the communication between "Auto" and the user. "Auto" either allows the user to input his own code or to enter the information it needs to generate the code.

Figure 4 is the final code "Auto Programming" has generated. It is important to notice that the final code uses the user's variable names and their correct size, and that the labels generated by "Auto" do not conflict with that generated by the user.

B. Conclusion

The "Auto-Programming" package offers the following three advantages:

(1) Ease of Operation: The software package is self-explanatory. The user needs to know only how to gain access to the system. Once a connection is established, the system will ask the user for all the information needed to generate the correct program.

(2) Reliability: The code stored in the library will be pre-tested and debugged.

(3) Time and Cost: All indications are that both time and cost of writing and debugging programs are reduced. An attempt will be made to substantiate the above statement. The example discussed in this report required 10 minutes of terminal time and 18 seconds of C.P.U.

Rome Air Development Center
F30602-74-C-0294

E. Lipshitz, M. L. Shooman and H. Ruston

```
WHAT BLOCK DO YOU WANT?
INPUT 1 FOR CONTROL BLOCK
INPUT 2 FOR A FUNCTION BLOCK
INPUT 3 FOR STOP BLOCK
INPUT 4 FOR OTHER BLOCK
INPUT 5 FOR END BLOCK
INPUT 6 FOR AN ERROR
```

```
4
1
4
2
3
4
2
4
1
4
3
4
3
```

FLOW CHART OF PROGRAM

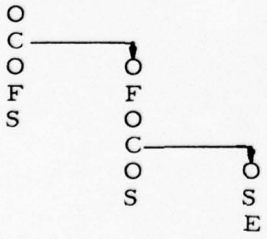


Fig. 2. Flow chart generation.

```
EACH LINE ENDS WITH *
LAST LINE ENDS WITH +
       INTEGER CLAS (10)*
       INTEGER GRAD (10)*
       READ (5, 3) J*
     3 FORMAT (I1)+
EACH LINE ENDS WITH *
LAST LINE ENDS WITH +
       IF(J.EQ.1) GO TO 7+
EACH LINE ENDS WITH *
LAST LINE ENDS WITH +
       DO 2 I=1, 10*
     2 READ (5, 5) CLAS (I)*
     5 FORMAT (A4)+
INPUT 10 FOR SORT
INPUT 11 FOR SEARCH
10
WHICH SORT DO YOU WANT
INPUT 13 FOR INTERCHANGE SORT
INPUT 12 FOR SHELL SORT
INPUT 14 FOR BUCKET SORT
12
INPUT TABLE NAME, SIZE
CLAS 10
EACH LINE ENDS WITH *
LAST LINE ENDS WITH +
     7 READ (5, 8) NAME, IGRD*
     8 FORMAT (A4, I3)+
INPUT 10 FOR SORT
INPUT 11 FOR SEARCH
11
INPUT 15 FOR LINEAR SEARCH
INPUT 16 FOR BINARY SEARCH
15
INPUT TABLE NAME SIZE AND WORD BEING SEARCHED FOR, LOCATION INDEX
CLAS 10NAMEK
EACH LINE ENDS WITH *
LAST LINE ENDS WITH +
       GRAD(K)=IGRD*
       READ(5, 9)J*
     9 FORMAT(R1)+
EACH LINE ENDS WITH *
LAST LINE ENDS WITH +
       IF(J.EQ.121) GO TO 10+
EACH LINE ENDS WITH *
LAST LINE ENDS WITH +
       GO TO 7+
EACH LINE ENDS WITH *
LAST LINE ENDS WITH +
    10 DO 12 I=1, 10*
    12 WRITE (6, 15) CLAS(I), GRAD(I)*
    15 FORMAT(1H  , A4, 10X, I3)+
```

Fig. 3. Illustration of the communication between the program auto and the user.

```
      INTEGER CLAS(10)
      INTEGER GRAD(10)
      READ(5,3)J
    3 FORMAT (I1)
      IF(J.EQ.1) GO TO 7
      DO 2 I=1,10
    2 READ(5,5) CLAS(I)
    5 FORMAT(A4)
C     SHELL SORT
C     INITIALIZE
      N=10
      ID=N
    6 ID=(ID+1)/2
      I=1
      I1=ID+1
      IFLG=0
C     COMPARE, REPLACE, SET FLAG
   11 IF(CLAS(I).LT.CLAS(I1)) GO TO     1
      ITEMP=CLAS(I)
      CLAS(I)=CLAS(I1)
      CLAS(I1)=ITEMP
      IFLG=1
C     PICK UP NEXT PAIR IF NOT REACHED END OF TABLE
    1 I1=I1+1
      IF(I1.GT.N) GO TO     4
      I=I+1
      GO TO 11
C     IS TABLE SORTED
    4 IF(IFLG.GT.0) GO TO     6
      IF(ID.NE.1) GO TO     6
      STOP
    7 READ(5,8) NAME,IGRD
    8 FORMAT(A4,I3)
C     LINEAR SEARCH
      DO   16 I=1, 10
      IF(CLAS(I).EQ.NAME) GO TO   14
   16 CONTINUE
      WRITE (6,  13)
   13 FORMAT(1H ,6X,SEARCH FAIL)
      STOP
   14 K   =I
      GRAD(K)=IGRD
      READ(5,9)J
    9 FORMAT(R1)
      IF(J.EQ.121) GO TO 10
      GO TO 7
      STOP
   10 DO 12 I=1,10
   12 WRITE(6,15)CLAS(I),GRAD(I)
   15 FORMAT(1H ,A4,10X,I3)
      STOP
      END
STOP
```

Fig. 4. The final code.

IMPLEMENTATION OF SHOOMAN'S MODEL OF EXHAUSTIVE TESTING: AN AUTOMATIC TYPE 1.A TESTER

D. L. Baggi and M. L. Shooman

In an internal paper, "Analytical Models for Software Testings," Martin L. Shooman describes a scheme for implementing a driver program to automatically test each path of a given program. An implementation of this scheme in PL/1, with a few revisions, is described here, along with two examples of programs, which were run with normal analytical debugging techniques, i.e., with some testing data -- and through the testing program. Comparisons among man-hour efforts and computer time in both cases are made.

A. The Testing Driver Program

A program, referred to as driver program, has been developed and run in conjunction with two programming examples. Its purpose is to allow automatic testing of all possible paths of any given program. A description of its functioning follows.

The driver program requires a data card containing an integer, N-TESTS, i.e., the number of IF statements, plus the number of repetitive DO groups, in the program and subroutines, to be supplied by the programmer.

The next data item has to be an "order," i.e., a character string such as 'NORMAL OPERATION,' or 'TEST,' or any other string. If the order is 'NORMAL OPERATION,' then the driver allows normal functioning of the program to be tested, e.g., with a set of data designed, by the programmer, to test some cases -- a normal debugging practice.

If the order is 'TEST,' no data set is needed for the tested program, but an array, T, with lower bound 1 and upper bound N-TESTS (the number of tests, as read in previously), will be constructed to represent, in ascending order, all possible bit combinations of binary numbers from 0 up to 2** N-TESTS - 1; this array is called testing word, and it thus consists of the bits of a binary counter with N-TESTS bits. Notice that the value 0 is fact represented, in the corresponding T, by -1, while 1 is represented by 1. Eventually the program to test is run for any such binary combination.

For any other order string, such as 'ENOUGH,' as well as in the case of absence of data, the whole system stops.

B. The Tested Program

Although no particular care has been taken to make sure that the driver program is fully compatible with all possible programs to test it is believed that, at its present state, the invariant part is flexible enough to accept a large class of programs with no modification, requiring for other programs only minor, sensible changes.

Shooman indicates (in pg. 5-1 of his paper), a strategy for implementing the driver program, namely, a revised way of writing IF statements and DO loops in a program to be submitted to testing; however, since such schemes lack generality (i.e., only conditions of the type "exp > 0" are allowed in IF statements, and only limits from 1 up to an upper bound > 0 in DO loops), the scheme described here has been developed as a natural derivative of these suggestions; hence, the only restrictions to be obeyed in writing a program will be the following:

(1) a) instead of IF cond THEN $statement_1$; ELSE $statement_2$;
write IF F(cond) THEN $statement_1$; ELSE $statement_2$;
b) instead of DO I = LIMIT 1 TO LIMIT 2 BY INCR;
write DO I = GL (LIMIT 1, LIMIT 2) TO GH BY INCR;
c) instead of DO WHILE (cond);
write DO WHILE (H(cond));
(where: F, GL, GH and H are described in a forthcoming technical report.)

(2) function and subroutine procedures are possible but should be internal to the program

(3) variables used in the program which are assigned a value through a read(GET LIST) statement should be initialized, for instance through a DCL INIT statement.

All these restrictions could be removed. To remove (1), one could construct a subprogram in the operating system which automatically supplies F, GL and GH, and H. Subroutines could be external, as long as a mechanism is provided for passing back and forth T and K, e.g., COMMON statements in FORTRAN, hence removing (2). Point (3) is a direct consequence of the read-in scheme described in I, which could be modified at will.

C. The Two Examples

For illustration two programs were chosen at the two ends of a spectrum: one with many IF statements, and input data, and the other with DO groups and one subroutine, and no input data.

1. First Example: Computer Solution of a Card Game

This is a very slightly altered version of Shooman's algorithm of Fig. 4-3 in his paper. The algorithm appears in Figure 1 here. It determines the winner of a card game, in which player A is dealt two cards, A1 and A2, and player B, similarly, gets B1 and B2, (four integers read in with a GET LIST statement). If both winners have a pair, the highest pair wins, or if they are equal it is a tie; if only one player has a pair, he wins; otherwise, the highest card wins, or if they are equal, the highest second card wins; identical hands are ties. At first the system was run under 'NORMAL OPERATION' conditions. The results are shown in a forthcoming report. The system was

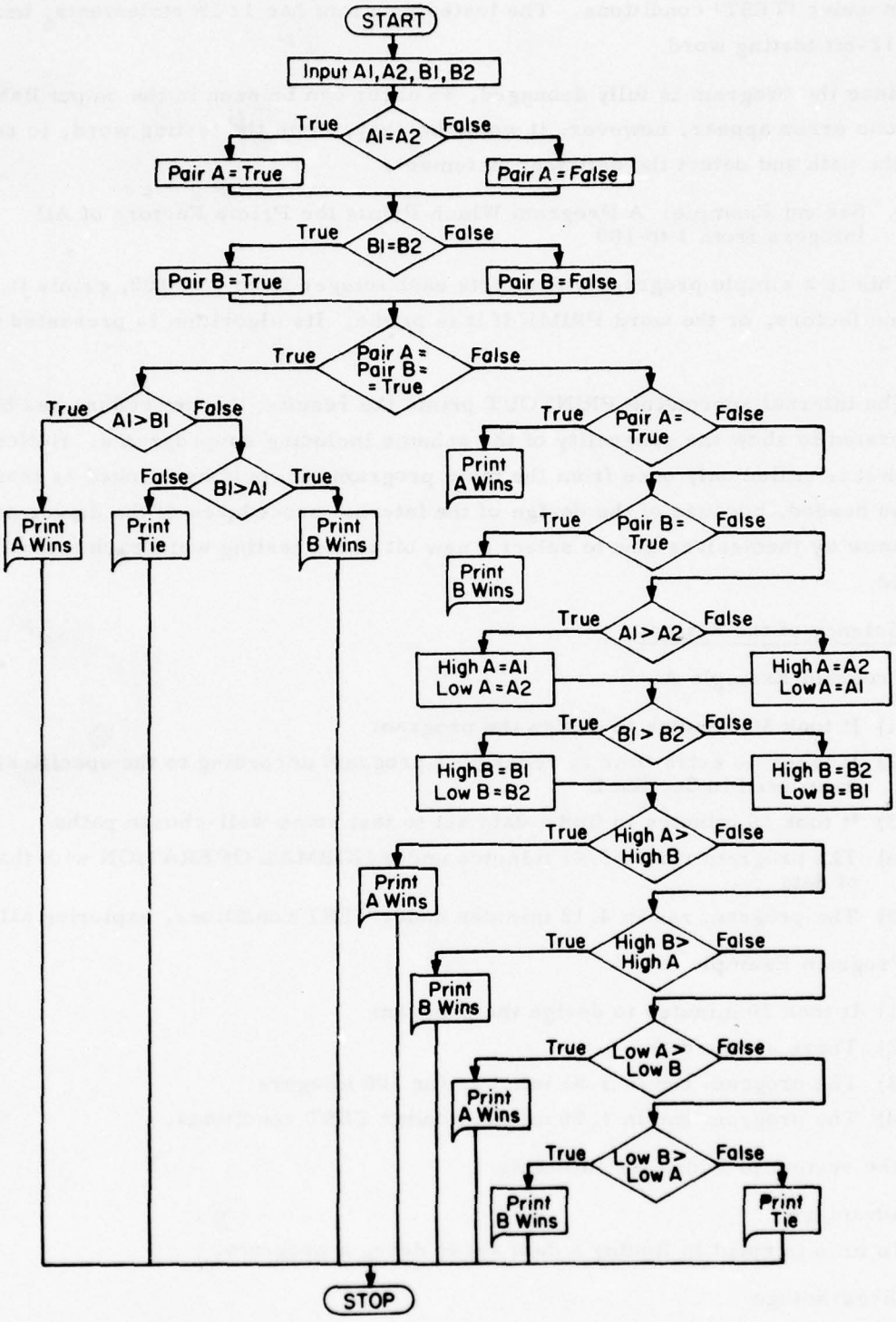


Fig. 1. Flow chart for computer solution of a card game.

next run under 'TEST' conditions. The tested program has 12 IF statements, hence we have a 12-bit testing word.

Since the program is fully debugged, no error can be seen in the output listing; should one error appear, however, it would be easy, from the testing word, to reconstruct the path and detect the deficient statement.

2. Second Example: A Program Which Prints the Prime Factors of All Integers from 1 to 100

This is a simple program which tests each integer from 1 to 100, prints it, and its prime factors, or the word PRIME if it is prime. Its algorithm is presented in Figure 2.

The internal subroutine PRINTOUT prints the results; this procedure has been incorporated to show the generality of the scheme including subprograms. Notice that, although it is called only once from the main program, it could be invoked as many times as needed, because of the design of the internal procedures of the driver program, which know by themselves how to select a new bit in the testing word each time they are used.

D. Efficiency of the System

Program Example A

(1) It took 30 minutes to design the program

(2) It takes no extra time to redesign a program according to the specifications expressed in Section B

(3) It took 10 minutes to find a data set to test some well-chosen paths

(4) The program ran in 2.41 minutes under NORMAL OPERATION with that set of data

(5) The program ran in 4.12 minutes under TEST conditions, exploring all paths.

Program Example B

(1) It took 20 minutes to design the program

(2) There are no data

(3) The program ran in 3.81 minutes for 100 integers

(4) The program ran in 1.76 minutes under TEST conditions.

Hence the system provides the following:

Advantage

No time is spent in finding a data set to debug a program.

Disadvantage

Running time may increase exponentially with the number of IF statements and DO loops. For instance, Program A contains 12 IF statements, and therefore the

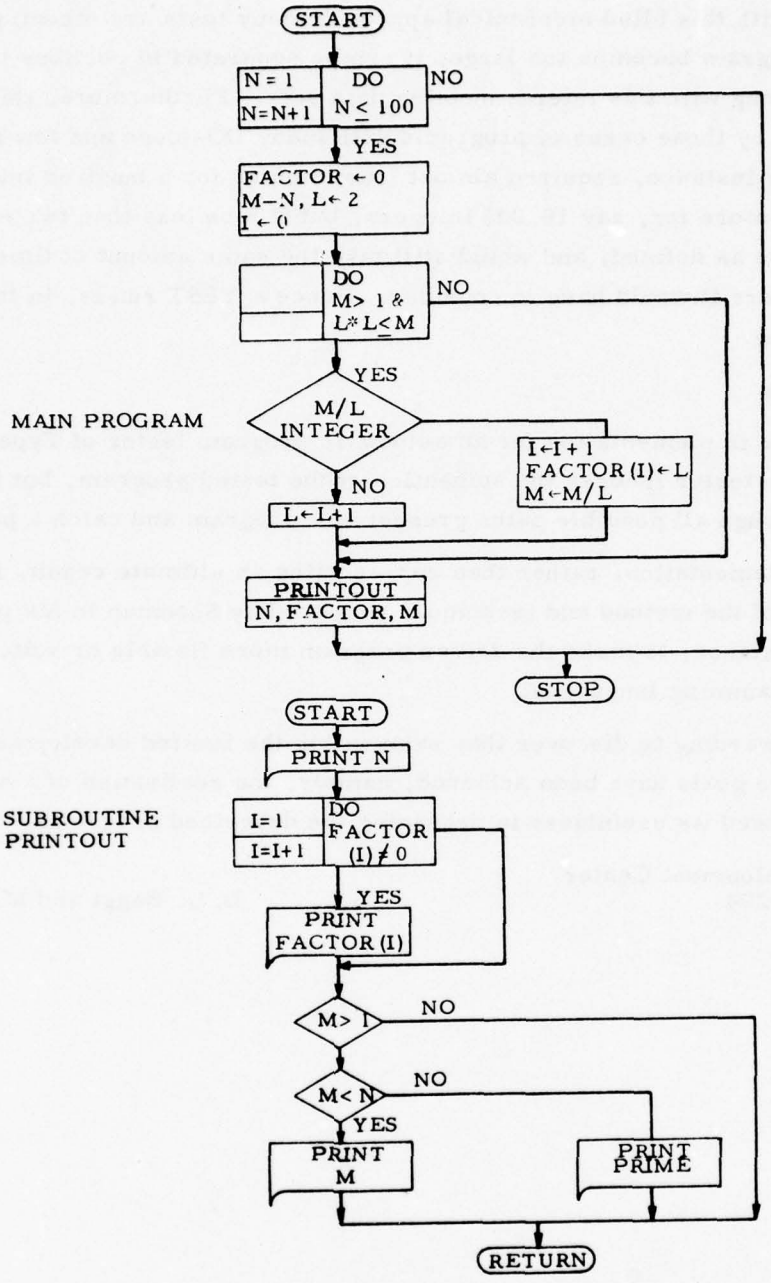


Fig. 2. Algorithm for printing prime factors from 1 to 100.

testing word has 13 bits and 8192 runs through the program were needed to test its 100 paths (hence, with this blind mechanical approach many tests are meaningless). However, if the program becomes too large, it can be separated in portions to be tested individually, along with this interconnecting data sets. Furthermore, this disadvantage is compensated by those cases of programs with many DO-loops and few IF statements. Program B, for instance, required almost four minutes for a hundred integers, and would use a lot more for, say 10,000 integers, but it took less than two minutes to go through all paths as defined, and would still take the same amount of time no matter how many integers it would have to consider. Hence a TEST run is, in these cases, very time saving.

E. Conclusions

A possible implementation for an automatic program tester of Type 1.A has been discussed. The tester ignores the semantics of the tested program, but is able however to run through all possible paths present in a program and catch a possible error.

This implementation, rather than representing an ultimate result, it meant to be an illustration of the method and techniques proposed by Shooman in his paper; it would be easy, for instance, to make the driver program more flexible or suited to other styles or programming languages.

It was rewarding to discover that even within the limited development of these techniques some goals have been achieved, namely, the realization of a system which has already proved its usefulness in debugging the described programs.

Rome Air Development Center
F30602-74-C-0294

D. L. Baggi and M. L. Shooman

OPTIMAL INSPECTION SCHEDULES FOR FAILURE DETECTION WHEN TESTS HASTEN FAILURES

L. Shaw and N. Wattanapanom

Barlow, Hunter and Proschan[1,2,3] posed and solved an interesting class of inspection problems related to reliability and maintainability of systems. In their models, a system operates for random time until it fails, but the failure can be detected only after a test costing c_1 has been performed. For example, judgment of whether or not a radar system's detection and false-alarm probabilities are acceptable might require connection of special test equipment and injection of special testing signals.

One approach to this problem seeks optimal testing times to minimize a loss function of the form

$$L = E\{c_1 N + c_2 d\} \tag{1}$$

in which N is the number of tests until the first one after the failure, and d is the time between the failure time and the subsequent detection time t_N. Reference 3 gives algorithms for finding the optimal inspection times t_i for lifetime distributions which are uniform, exponential or Polya of order 2, when using the single-cycle (no renewal) loss function in Equation (1).

A second point of view postulates repair or replacement after each failure detection, with associated average cost and average repair time requirements. In this case the performance is judged using the mean cost per unit time or equivalently[4] the ratio of mean-cost-per-renewal to mean-time-between-renewals. Reference 3 shows how techniques used for the single-cycle problem can be modified to find the best testing times for the case with renewals.

Here we seek similar solutions but include the possibility that the mechanical and electrical stresses of the test might destroy or damage the system. Mathematically, we allow the ith test to destroy the system with probability β or, with probability $(1 - \beta)$, to increase the failure rate (reciprocal of mean time-to-failure) to $\lambda_i > \lambda_{i-1}$ without changing the form of the conditional lifetime distribution. These generalizations are approached in a direct manner by using dynamic programming to set up recursive optimization equations. The main contribution here is the presentation of convergent algorithms for solving the optimization equations. Sensitivity of inspection policy performance to uncertainty in failure rate and degradation parameters is also studied.

The initial phases of this work, dealing with single cycle (no renewal) cases for exponential and uniform conditional lifetime distributions, were summarized in last year's research summary.[5] Here we show how those results generalize to the case in

which the failed part is replaced after each failure detection, and the lifetime distributions are exponential. More details about this case, as well as those for other conditional distributions can be found in Reference 6.

When the mean cost for component replacement and mean time to accomplish that replacement are denoted by s and r, respectively, then the ith cycle, starting at time t_{0_i} and ending at t_{N_i}, will accrue a mean cost

$$C_i = N_i c_1 + d_i c_2 + s \tag{2}$$

and will have a duration

$$\tau_i = (t_{N_i} - t_{0_i}) + r \tag{3}$$

Using these terms the average cost is defined as

$$R(\delta) \triangleq \lim_{K \to \infty} \frac{\sum_{i=1}^{k} C_i}{\sum_{i=1}^{k} \tau_i} \tag{4}$$

The argument δ is shown to emphasize the dependence of this average cost on an inspections policy $\delta = \{\delta_0, \delta_1, \cdots\}$.

It is well known[4] that $R(\delta)$ can be expressed in terms of the mean cost per cycle $\overline{C}$ and the mean cycle duration $\overline{\tau}$, as

$$R(\delta) = \overline{C}/\overline{\tau} \tag{5}$$

Both the numerator and denominator of Eq. (5) are affected by the inspection policy δ. However, it is possible to find the δ which minimizes $R(\delta)$ by considering an auxiliary optimization problem having the cost function

$$\mathcal{L}(\mu, \delta) = \overline{C} - \mu\overline{\tau} \tag{6}$$

These two problems are related as follows.

Theorem:[3] If there exists a $\mu = \mu^*$ for which $\min_{\delta} \mathcal{L}(\mu^*, \delta) \triangleq \mathcal{L}(\mu^*, \delta(\mu^*)) = 0$, then the schedule $\delta(\mu^*)$ also minimizes $R(\delta)$.

The existence of such a desired μ^* is quite evident for the present problem. Equation (6) can be rewritten as

$$L(\mu, \delta) = c_1\overline{N} + c_2\overline{d} + s - \mu[\overline{\tau} + r] \tag{7}$$

a little thought shows that when $\mu \geq c_2$, increases in δ_i will decrease $\mathcal{L}$ since $\overline{\tau} > \overline{d}$.

Thus, for $\mu \geq c_2$ the optimal policy is to make no inspections or repairs, with a corresponding optimal $\mathcal{L} < 0$. On the other hand, when $\mu = 0$, $\mathcal{L} > 0$ for any policy -- including the optimal one. We conclude that $\mathcal{L}(\mu, \delta, (\mu))$ changes from a positive value at $\mu = 0$ to a negative one for $\mu = c_2$, and this change must be a continuous one passing through zero at least once in view of the smooth nature of all densities and cost terms entering into the loss functions.

An algorithm which uses the Theorem to optimize inspection time in the presence of renewals consists of two parts. An algorithm like the single-cycle one in Ref. 5 is needed to minimize Eq. (7) for each choice of μ until $\mathcal{L}(\mu, \delta(\mu)) = 0$. The first part parallels the approach in the single-cycle case since $\mathcal{L}$ defined in Eq. (7) is a single cycle loss with μ similar to c_3.

In particular, Eq. (7) leads to the following recursive expression for the minimal mean future loss after test κ,[6]

$$L_\kappa^o = \min_{\delta_\kappa} \{c_1 + (c_2 - \mu)\delta_\kappa - c_2/\lambda_\kappa + (s - \mu r) + e^{-\lambda_\kappa \delta_\kappa}(L_{\kappa+1}^o - s + \mu r + c_2/\lambda_\kappa)\} \tag{8}$$

Differentiation yields the recursive optimization equations.

$$\begin{aligned} &\text{(i)} \quad \delta_\kappa^o = \lambda_\kappa^{-1} \ell n[(c_2 + \lambda_\kappa L_{\kappa+1}^o - \lambda_\kappa(s - \mu r))/(c_2 - \mu)] \\ &\text{(ii)} \quad \mathcal{L}_\kappa^o = c_1 - \mu/\lambda_\kappa + (s - \mu r) + (c_2 - \mu)\delta_\kappa^o \end{aligned} \tag{9}$$

It can be shown[6] that the $\mu < c_2$ condition derived above insures that the zero-derivative condition corresponds to a minimum and that all $\delta_\kappa^o > 0$. Finally, the Lemma in the single cycle case must be modified for the present loss function to

$$\underline{\text{Lemma}}: \lim_{\kappa \to \infty} \mathcal{L}_\kappa^o = c_1 - \mu r + s \tag{10}$$

Once the μ^* of the Theorem has been found, it follows immediately that

$$\min_\delta R(\delta) = \mu^* \tag{11}$$

Sensitivity

The preceding paragraphs describe how to get the best inspection schedule δ and the corresponding minimal loss per unit time. It is also of interest to see how sensitive this solution procedure is to the precision of knowledge of model parameters, e.g., λ_0, and how sensitive it is to the accuracy of the iterative calculations. It is straightforward to use the exponential conditional densities to get iterative expressions for $\overline{N}$, $\overline{d}$ and $\overline{\tau}$ which can be terminated when the additional terms contribute very little.[6]

Examples

Tables I, II and III show results for numerical examples based on the models in this section. Table I shows the small sensitivity of $R(\delta)$ to use of non-optimal testing schedules that result when $\mu \neq \mu^*$. The parameters for Tables II and III differ only in that the latter has non-zero values for r and s, the renewal costs. Both of those tables show minimum R values for several λ_0 as well as the performance degradation when $\lambda_0 = 5$ but, when the optimum δ-schedule for a different (mismatched) λ_0 value is used.

TABLE I. Mean loss rate minimization - exponential case

$c_1 = 1.0$, $c_2 = 20.0$, $s = 0$, $r = 0$, $\lambda_0 = 5.0$, $M = 21$, $\rho = 0.9$, $\kappa = 20$, and $\lambda_K = \lambda_0/\rho^K$

μ	R
10	12.72591
11	12.67058
12	12.63822
$\mu^* = 12.63183$	12.63200
13	12.63429
14	12.66670
15	12.74756
16	12.89658
17	13.14914
18	13.57903
19	14.39155

TABLE II. Minimum loss rates - exponential case; sensitivity to λ_0 error

$c_1 = 1.0$, $c_2 = 20$, $s = 0$, $r = 0$, $\rho = 0.9$, $M = 21$, and $\lambda_K = \lambda_0/\rho^K$

λ_0	$R(5)_{mismatch}$	R_{min}
2	13.05727	8.68520
3	12.75710	10.27669
4	12.65410	11.55473
5	12.63183	12.63200
6	12.64438	13.56575
7	12.66983	14.38974
8	12.69736	15.12581
9	12.72094	15.78896
10	12.73731	16.38998

TABLE III. Mean loss rates - exponential case; sensitivity to λ_0 error - non zero renewal costs.

$c_1 = 1.0$, $c_2 = 20.0$, $s = 1.2$, $r = .001$, $\rho = 0.9$, $M = 21$, and $\lambda_K = \lambda_0/\rho^K$

λ_0	$R^{(5)}_{mismatch}$	R_{min}
2	16.28924	10.62738
3	16.22188	12.87934
4	16.21368	14.70045
5	16.21360	16.21360
6	16.24733	17.47788
7	16.38830	18.52204
8	16.38830	19.35293

Office of Naval Research
N00014-75-C-0858

L. Shaw and N. Wattanapanom

National Science Foundation
ENG 73-03802

Joint Services Technical Advisory Committee
F44620-74-C-0056

REFERENCES

1. R.E. Barlow and L. Hunter, "Mathematical Models for System Reliability, Part 2," The Sylvania Technologist 13, 2, 60-65 (April 1960).

2. R.W. Barlow, L. Hunter and F. Proschan, "Optimum Checking Procedures," J. Soc. Industrial Math., 11, 4, 1078-1095 (December 1963).

3. R.E. Barlow and F. Proschan, "Mathematical Theory of Reliability," Chap. 4 (Wiley, 1965).

4. S. Ross, "Applied Probability with Optimization Applications," (Holden Day, 1970).

5. N. Wattanapanom and L. Shaw, "Optimal Inspection Schedules When Tests Hasten Failures," Progress Report No. 40 to JSTAC, Polytech. Inst. of New York, Report No. R-452.40-75, 579 (1975).

6. N. Wattanapanom, "Optimal Inspection Schedules for Failure Detection When Tests Hasten Failure," Ph.D. Dissertation (Operations Research), Polytech. Inst. of New York (June 1976). (Summarized in Research Report POLY EE/EP 76-001, (January 1976).)

RELIABILITY APPLICATIONS OF A MULTIVARIATE EXPONENTIAL DISTRIBUTION

L. Shaw and C-L. Hsu

Several types of multivariate exponential distributions have been proposed for various reliability applications.[1,2] These are joint distributions for which the univariate marginal distributions are negative exponentials. The present work has examined one particular class of these distributions for both modeling of downtime distributions and modeling of stages of component deterioration.

The n non-independent exponential random variables

$$r_1, r_2, \cdots, r_n \tag{1}$$

can be defined with respect to two sets of normal random variables. In particular, let $\underline{w}$ and $\underline{z}$ be two stochastically independent n-vectors, each being normally distributed with a zero mean-value vector, and each with the same covariance matrix Γ. It is easy to show[3] that defining

$$r_i = w_i^2 + z_i^2 \tag{2}$$

makes the r_i correlated exponential random variables with mean values

$$E[r_i] = 2\gamma_{ii} \tag{3}$$

and correlation coefficients

$$\rho_{ij} = \gamma_{ij}^2 \tag{4}$$

The precise form of the corresponding multivariate density will depend on the type of Γ matrix selected. For example, if

$$\gamma_{ij} = \sqrt{\sigma_i^2 \sigma_j^2}\, \rho^{|i-j|} \tag{5}$$

then the bivariate and trivariate densities are

$$f(r_1, r_2) = \left[4\sigma_1^2 \sigma_2^2 (1-\rho^2)\right]^{-1} \exp\left[-\frac{r_1/\sigma_1 + r_2/\sigma_2}{2(1-\rho^2)}\right] I_0\left[\frac{\sqrt{\rho^2 r_1 r_2}}{\sigma_1 \sigma_2 (1-\rho^2)}\right] \tag{6}$$

$$f(r_1, r_2, r_3) = \frac{1}{8\sigma_1^2 \sigma_2^2 \sigma_3^2 (1-\rho^2)^2}\, e^{-\left[\left(\frac{r_1}{\sigma_1^2} + \frac{r_2(1+\rho^2)}{\sigma_2^2} + \frac{r_3}{\sigma_3^2}\right)/2(1-\rho^2)\right]}$$

$$\cdot\, I_0\left[\frac{\sqrt{r_1 r_2 \rho^2}}{\sigma_1 \sigma_2 (1-\rho^2)}\right] I_0\left[\frac{\sqrt{r_2 r_3 \rho^2}}{\sigma_2 \sigma_3 (1-\rho^2)}\right] \tag{7}$$

in which I_0 is the modified Bessel Function of order zero.

A. Modeling

The design of maintenance policies for maintainable and repairable systems makes use of information about probability distributions of component lifetimes and downtimes. Since such designs are facilitated if these distributions have simple analytic forms, downtime distributions have been frequently modeled as lognormal, Weibull or Erlarg in form.[4] These distributions all are skewed and correspond to non-negative random variables like downtimes. Here we consider one more family of distributions which has some physical motivation.

Since a downtime interval is often the sum of subsidiary intervals (for failure isolation, component removal, repair, reassembly, alignment, etc.) it seems reasonable to think of the downtime x_n as a sum of subsidiary time intervals

$$x_n = \sum_{i=1}^{n} r_i \tag{8}$$

The subscript on x_n reminds us of the number of summands. Several distributions are possible for the individual r_i, but here we consider exponential distributions which are the simplest and which are widely used to represent random times between events.

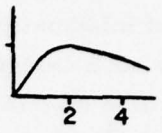
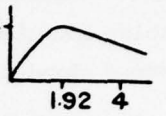
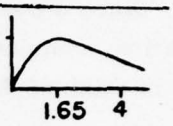
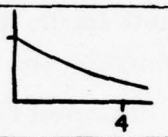
It is well known that the downtime x_n will have an Erlang distribution if the r_i are independent exponential variables with identical mean values. Muth[5] has considered the approximation of Weibull and lognormal distributions by x_n in which the r_i are independent exponential variables but with possibly different mean values. Here we further generalize to allow dependence among the r_i -- a reasonable situation if the variables represent related steps in a sequence of downtime operations. In particular, we use r_i with the multivariate exponential distribution described above.

Figure 1 shows some possible densities for x_2 (i.e., when n = 2 in Equation (8)). Each density shown there can be developed in two different ways. One approach assumes r_1 and r_2 are independent, with unequal mean values, and $f(x_2)$ can be found by convolution of their marginal exponential densities. Alternatively, when the r_i are not independent, they can be expressed in terms of their normally distributed generators as

$$x_2 = w_1^2 + w_2^2 + z_1^2 + z_2^2 \tag{9}$$

The characteristic function

$$\Phi_{x_2}(s) = E\left[e^{-s(w_1^2 + w_2^2 + z_1^2 + z_2^2)}\right] \tag{10}$$

(a) $E[r_1] = E[r_2] = 2$	(b) $\rho = 0$		
i) $\rho = 0$	$E[\hat{r}_1] = 2$ $E[\hat{r}_2] = 2$	$f_{\underset{\sim}{x}_2}(x_2) = \frac{1}{4} x_2 e^{-0,5 x_2}$	
ii) $\rho = \frac{1}{4}$	$E[\hat{r}_1] = 2.5$ $E[\hat{r}_2] = 1.5$	$f_{\underset{\sim}{x}_2}(x_2) = e^{-0,4 x_2} - e^{-\frac{2}{3} x_2}$	
iii) $\rho = \frac{1}{2}$	$E[\hat{r}_1] = 3$ $E[\hat{r}_2] = 1$	$f_{\underset{\sim}{x}_2}(x_2) = \frac{1}{2}\left(e^{-\frac{1}{3} x_2} - e^{-x_2}\right)$	
iv) $\rho = \frac{3}{4}$	$E[\hat{r}_1] = 3.5$ $E[\hat{r}_2] = 0.5$	$f_{\underset{\sim}{x}_2}(x_2) = \frac{1}{3}\left(e^{-\frac{2}{7} x_2} - e^{-2 x_2}\right)$	
v) $\rho = 1$	$E[\hat{r}_1] = 4$ $E[\hat{r}_2] = 0$	$f_{\underset{\sim}{x}_2}(x_2) = \frac{1}{4} e^{-\frac{1}{4} x_2}$	

(a) Equal mean and correlated
(b) Independent with unequal means

Fig. 1. Densities for sums of exponential variables.

can be computed directly using the properties of w_i and z_i. For example, when $\gamma_{ii} = 1$ and $\rho_{w_1 w_2} = \rho$ then

$$\Phi_{x_2}(s) = [4(1 - \rho^2) s^2 + 4 s + 1]^{-1} \tag{11}$$

and the densities in Fig. 1 can be found by Laplace inversion for various values of ρ.

The equivalence of a sum of independent exponential random variables to the sum of different, but correlated, exponential random variables suggested the following.

Theorem: All possible density functions for x_n defined in Eq. (2) and Eq. (8) can be achieved using independent r_i, i.e., with diagonal $\Gamma = \mathrm{diag}(a_1, a_2, \cdots, a_n)$.

This theorem is proved in Ref. 3 by showing that with $\Gamma = MM^T$ then

$$\Phi_{x_n}(s) = (|I + 2 s M^T M|)^{-1} = \frac{1}{q(s)} \tag{12}$$

It can be argued that q(s) is an n-th degree polynomial with real, negative roots and q(0) = 1. Thus, given Eq. (12) for any arbitrary Γ, some other diagonal Γ_D can be found to produce the same q(s) and $\Phi_{x_n}(s)$.

In general, the n-independent exponential variables whose sum is indistinguishable from the sum of n-correlated ones will have different mean values from those of the correlated variables.

Once this structure has been established for sums of correlated exponential variables, previous results for sums of independent variables (e.g., those of Muth[5]) are directly applicable. However, it appears that when equal mean variables have their mean and correlation ρ in Eq. (11) adjusted so x_n matches the lognormal mean and variance, then f_{x_n} tends to lognormal for large n. This is in contrast to Muth's summing of independent variables in which he had to search for the proper individual mean values which made the sum approximately lognormal.

B. Deterioration Modeling for Replacement Schedule Optimization

One model frequently used to describe component deterioration assumes that a new component retains its good qualities for a random time duration, after which it performs for another random time interval with reduced quality, etc. The quality k is thus assumed to take on a finite set of values, say $k = 0, 1, \cdots, N$, with k = 0 for a new part and k = N for a worthless part. Luss[6] has developed optimal inspection schedules (for observing k) and replacement rules (for resetting k = 0) with respect to a specified reward measure. We are attempting to generalize his work with independent exponential durations in each stage, by using the correlated exponential variables, defined above, to represent the durations of the stages.

In the simplest case we have assumed that a system is characterized by a single deterioration state k; transitions from k = i to i + 1 are immediately observable; replacement (setting k = 0) is mandatory when k = N; rewards are decreasing functions of k; and the desired replacement rule must maximize the average reward per unit time.

Two kinds of reward structures were considered for a system in state k = i:

1. Linear: c_i dollars per unit time
   $c_0 > c_1 > c_2 \cdots c_N$
2. Quadratic: $c_i t$ per unit time after t seconds in state i
   $c_0 > c_1 > \cdots > c_N$

In the linear case the optimal replacement rule is to replace when k reaches some k^*, independent of $r_0, r_1, \cdots$, the durations in states $k = 0, k = 1, \cdots$; and independent of the correlations between those durations.

In the quadratic case, dynamic programming arguments have shown that the best replacement rule is of the form: replace on entering k if $r_{k-1} < r^*_{k-1}$. This rule reflects the positive correlation between successive durations. A small r_{k-1} yields predictive information that the rate of deterioration is great. It is anticipated that future work will show that the decision thresholds are ordered according to

$$r^*_{k-1} < r^*_k$$

Derivation of the quadratic case results required study of properties of the conditional densities of the r_i sequence. In particular, the r_i sequence preserves the Markov property of the underlying gaussian sequences defined by Eq. (5); and the conditional variables are stochastically increasing in the sense that the conditional distribution function $f(r_k/r_{k-1})$ is an increasing function of r_{k-1} for every r_k.

Office of Naval Research
N00014-75-C-0858

National Science Foundation
ENG 73-03802

L. Shaw and C-L. Hsu

REFERENCES

1. R.E. Barlow and F. Proschan, "Statistical Theory of Reliability and Life Testing," Hold, Rinehart and Winston (1975).
2. A.W. Marshall and I. Olkin, "A Multivariate Exponential Distribution," J. of the American Statistical Assoc., 62, 30-44 (1967).
3. C-L. Hsu and L. Shaw, "Downtime Distributions Based on a Multivariate Exponential Distribution," Tech. Report, POLY EE/EP 76-002 (February 1976).
4. A.S. Goldman and T.B. Slattery, "Maintainability," John Wiley and Sone (1964).
5. E.J. Muth, "Stochastic Theory of Repairable Systems," Ph.D. Dissertation (Systems), Polytech. Inst. of Brooklyn (June 1967).
6. H. Luss, "Maintenance Policies When Deterioration Can be Observed by Inspections," Operations Research, 2, 359 (1976).

NEW RESULTS IN SYSTEM THEORY: BAUER-TYPE FACTORIZATION OF POSITIVE MATRICES AND THE THEORY OF MATRIX POLYNOMIALS ORTHOGONAL ON THE UNIT CIRCLE

D.C. Youla and N.N. Kazanjian

A. Introduction

Some of the most impressive accomplishments in modern system theory hinge on the solution of the following problem. Consider an n x n rational matrix G(s) of the complex variable $s = \sigma + j\omega$ meeting the conditions*

(a) $G(s) = G_*(s)$;

(b) $G(j\omega) \geq 0$, all real ω ;

(c) $\det G(s) \not\equiv 0$.

Produce an effective construction for the n x n rational matrix W(s) which is analytic together with its inverse in Re $s > 0$ and satisfies

$$G(s) = W_*(s)\, W(s) \quad . \tag{1}$$

Clearly, for any such factor W(s), $\det W(s) \not\equiv 0$, Re $s > 0$, and W(s) is said to be minimum-phase. Many frequency-domain solutions to this problem have appeared since 1958 and are extensively discussed in the literature.[1-8]

Let

$$G(s) = \frac{G_0 + G_1 s + \ldots + G_m s^m}{g_0 + g_1 s + \ldots + g_n s^n} \equiv \frac{P(s)}{g(s)} \tag{2}$$

where the G_r's are n x n constant matrices, $r = 0 \rightarrow m$, and g(s) is the least common multiple of all denominators in G(s). Then, $g(s) = g_*(s)$, $P(s) = P_*(s)$, $g(j\omega) \geq 0$ and $P(j\omega) \geq 0$, ω real. As a consequence, $m = 2\nu$ and $n = 2\mu$ are both nonnegative even integers. Using the mapping

$$z = \frac{1 - s}{1 + s} \quad , \tag{3}$$

the closed right-half s-plane is imaged into the unit circle $|z| \leq 1$ and G(s) is transformed into

---

*The complex conjugate, transpose, adjoint, determinant and trace of a matrix A are denoted by $\bar{A}$, A', $A^*(\equiv \bar{A}')$, det A and Tr A, respectively; 1_n is the identity and 0_n the n x n zero matrix. If $A = A^*$, A is hermitian and if $A^*A = 1_n$, A is unitary. By $A - B \geq 0$ we mean that A-B is hermitian nonnegative-definite while $A - B > 0$ signifies that it is positive-definite. For a rational matrix A(s), $A_*(s) \equiv A^*(\bar{s})$. Evidently, if A(s) is real for real s, $A_*(s) = A'(-s)$ and always $A_*(j\omega) = A^*(j\omega)$, ω real. If $A(s) = A_*(s)$, A(s) is paraconjugate hermitian and if $A_*(s)\, A(s) = 1_n$, it is paraconjugate unitary. (In the special case that A(s) is real A(s) is said to be parahermitian or paraunitary.)

$$R(z) = G\left(\frac{1-z}{1+z}\right) = (1+z)^{n-m} \cdot \frac{N(z)}{d(z)} \quad , \tag{4}$$

N(z) and d(z) polynomial. Condition (a) forces

(a_1) $R(z) = R_*(z)$

where now $A_*(z) \equiv A^*(1/\bar{z})$, (b) requires

(b_1) $R(e^{-j\theta}) \geq 0$, all real θ

and of course

(c_1) $\det R(z) \not\equiv 0$.

In terms of

$$K(z) = \frac{N(z)}{z^{\nu}} \tag{5}$$

and

$$k(z) = \frac{d(z)}{z^{\mu}} \tag{6}$$

we can write

$$R(z) = \frac{K(z)}{k(z)} \cdot ((1+z)_* \cdot (1+z))^{\mu-\nu} \tag{7}$$

and (a_1), (b_1), (c_1) translate into

$$(a_2) \quad K(z) = K_*(z) \; ; \; K(e^{-j\theta}) \geq 0 \tag{8}$$

$$(b_2) \quad k(z) = k_*(z) \; ; \; k(e^{-j\theta}) \geq 0 \tag{9}$$

$$(c_2) \quad k(z) \cdot \det K(z) \not\equiv 0 \tag{10}$$

Let

$$K(z) = B_*(z)\, B(z) \quad , \tag{11}$$

and

$$k(z) = b_*(z)\, b(z) \tag{12}$$

where

$$B(z) = B_0 + B_1 z + \ldots + B_{\nu} z^{\nu} \quad , \tag{13}$$

$$b(z) = b_0 + b_1 z + \ldots + b_{\mu} z^{\mu} \tag{14}$$

and

$$b(z) \cdot \det B(z) \neq 0 \ , \ |z| < 1 \quad . \tag{15}$$

Then

$$R(z) = H_*(z)\, H(z) \tag{16}$$

in which

$$H(z) = \frac{B(z)}{b(z)} \cdot (1+z)^{\mu-\nu} \tag{17}$$

is obviously analytic together with its inverse in $|z| < 1$. Thus, transforming back,

$$W(s) = H\left(\frac{1-s}{1+s}\right) \tag{18}$$

and the construction of W(s) has therefore been made to depend on the availability of a procedure for minimum-phase polynomial factorization of self-inversive matrix polynomials in the variable z which are nonnegative-definite on the boundary of the unit circle.

At first glance, this reformulation of the problem in the z-domain appears to be of little significance since the slight advantage gained by working in the compact set $|z| \leq 1$ instead of Re $s \geq 0$ is more than offset by the need to carry out the transformation $G(s) \rightarrow R(z)$. Nevertheless, Bauer[9] in 1955 exploited the above correspondence in the scalar case to exhibit a novel technique well suited to computer implementation. Quite recently,[10] Kailath gave a partial generalization of Bauer's method to matrices by employing a functional-theoretic argument revolving around von Neumann's theorem on alternating projections.[11] However, his proof is valid only under the unnecessarily restrictive assumption $\det R(e^{-j\theta}) > m$, m a positive constant. Our objective in this report is to supply an elementary (and highly informative) proof of this matrical Bauer extension in its most general setting and to define and derive explicit formulas for a new class of associated matrix polynomials orthogonal on the unit circle. The intimate connection with the factorization problem is established by a detailed study of their limit properties.

Before embarking on section B we summarize for the reader's benefit several useful results of a purely mathematical nature.

Let A and B be two n x n hermitian nonnegative-definite matrices. Then,

$$\det(A+B) \geq \det A + \det B \quad . \tag{19}$$

In addition, if at least one of A and B, A say, is nonsingular, equality attains in Eq. (19) iff $B = \lambda A$, λ a nonnegative scalar $\neq 0$ for $n > 1$ (Reference 12). As an interesting

application, suppose $A - B \geq 0$ and $\det A = \det B \neq 0$. Clearly, from $A = (A - B) + B$ and Eq. (19) we conclude that

$$\det A \geq \det(A - B) + \det B = \det(A - B) + \det A \quad .$$

$$\therefore \det(A - B) = 0$$

and

$$A - B = \lambda B \text{ giving } \det A = (1 + \lambda)^n \cdot \det A \quad .$$

Hence $\lambda = 0$, $A = B$ and it is seen that under certain conditions the equality of the determinants of two matrices implies equality of the matrices. More generally, Eq. (19) is included in

$$\sqrt[n]{\det(A + B)} \geq \sqrt[n]{\det A} + \sqrt[n]{\det B} \quad , \tag{20}$$

Minkowski's inequality.[12]

The Master Inequality.[13,I] Let the n x n matrix $K(\theta)$ be hermitian nonnegative-definite and L_1; i.e., $K(\theta) \geq 0$ a.e. with entries absolutely integrable over $-\pi \leq \theta \leq \pi$. Then,

$$\ln \det \frac{1}{2\pi} \int_{-\pi}^{\pi} K(\theta)\, d\theta \geq \frac{1}{2\pi} \int_{-\pi}^{\pi} \ln \det K(\theta)\, d\theta \geq -\infty \quad . \tag{21}$$

Proof. Obviously, $K(\theta) \in L_1$ implies $\sqrt[n]{\det K(\theta)} \in L_1$ and since integration is a summation, Eq. (20) yields

$$\sqrt[n]{\det \frac{1}{2\pi} \int_{-\pi}^{\pi} K(\theta)\, d\theta} \geq \frac{1}{2\pi} \int_{-\pi}^{\pi} \sqrt[n]{\det K(\theta)}\, d\theta \quad . \tag{22}$$

Taking logarithms on both sides and using the inequality of the arithmetic and geometric means[14] we get the desired result Eq. (21), Q.E.D.

An n x n matrix function

$$F(z) = \sum_{r=0}^{\infty} F_r z^r \tag{23}$$

is said to be of class H_2 (Ref. 13,I) if

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \mathrm{Tr} \left\{ F^*(re^{-j\theta}) F(re^{-j\theta}) \right\} d\theta$$

is bounded for all $r < 1$. This requirement is equivalent to

$$\sum_{r=0}^{\infty} \mathrm{Tr}(F_r^* F_r) < \infty \quad . \tag{24}$$

Such a function is analytic in $|z| < 1$ and its radial interior limit

$$F(e^{-j\theta}) \equiv \lim_{r \to 1-0} F(re^{-j\theta}) \tag{25}$$

exists for almost all θ (Reference 15). The next inequality plays a central role in the sequel.

Jensen's Inequality.[15] Let $F(z) \in H_2$. Then, if $F(0)$ is nonsingular,

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \ln|\det F(e^{-j\theta})| \, d\theta \geq \ln|\det F(0)| = \ln|\det F_0| > -\infty \tag{26}$$

with equality iff $\det F(z)$ is devoid of zeros in $|z| < 1$. That is, iff $F(z)$ is minimum-phase.

Lastly, let $\{A_i\}$ be a sequence of $n \times n$ hermitian nonnegative-definite matrices which is either monotonically nondecreasing or monotonically nonincreasing and bounded above or below, respectively, by the $n \times n$ hermitian nonnegative-definite matrix H. Then, element by element,

$$A = \lim_{i \to \infty} A_i$$

exists as an $n \times n$ hermitian nonnegative-definite matrix and correspondingly,

$$A_i \leq A \leq H$$

or

$$A_i \geq A \geq H \quad ,$$

$i = 1 \to \infty$. The proof is carried out by perceiving that $A_i \leq A_{i+1} \leq H$, $i = 1 \to \infty$, is equivalent to asserting that the sequence $\{a^* A_i a\}$ is monotone nondecreasing and bounded above by $a^* H a$ for every choice of n-vector a.

B. The Main Results

Let the $n \times n$ matrix $K(\theta)$ belong to L_1 over $I = [-\pi \leq \theta \leq \pi]$. As is well known, such a matrix possesses a Fourier series

$$K(\theta) \sim \sum_{r=-\infty}^{\infty} A_r e^{jr\theta} \tag{27}$$

whose $n \times n$ matrix coefficients

$$A_r = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-jr\theta} K(\theta)\, d\theta \to 0_n \tag{28}$$

as $|r| \to \infty$ (Reference 16). Clearly, if $K(\theta)$ is hermitian, i.e., if $K(\theta) = K^*(\theta)$ a.e.,

$$A_{-r} = A_r^* \tag{29}$$

for all r and the block Toeplitz matrices

$$T_m = \begin{bmatrix} A_0 & A_1 & \cdot & \cdot & \cdot & A_m \\ A_{-1} & A_0 & \cdot & \cdot & \cdot & A_{m-1} \\ \cdot & \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & & \cdot & \cdot \\ A_{-m} & A_{-m+1} & \cdot & \cdot & \cdot & A_0 \end{bmatrix} \tag{30}$$

of respective sizes $(m+1)n \times (m+1)n$, $m = 0 \to \infty$, are obviously hermitian. Furthermore, if a.e. $K(\theta) = K(-\theta) =$ real matrix, all A_r are real symmetric matrices and $A_{-r} = A_r$.

Lemma 1. Suppose $K(\theta) \in L_1$. Then, $K(\theta)$ is hermitian nonnegative-definite, i.e., $K(\theta) = K^*(\theta) \geq 0$ a.e.,

(1) iff every T_m is nonnegative-definite, $m = 0 \to \infty$.

(2) Subject to the additional constraint

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \ln \det K(\theta)\, d\theta > -\infty \quad , \tag{31}$$

all T_m are positive-definite, $m = 0 \to \infty$.

Proof. Let $c_0, c_1, \ldots, c_m$ denote an arbitrary choice of m+1 constant n-dimensional column-vectors and put

$$f(\theta) = \sum_{r=0}^{m} c_r e^{-jr\theta} \quad . \tag{32}$$

Under the assumption that $K(\theta)$ is hermitian nonnegative-definite,

$$0 \leq \frac{1}{2\pi} \int_{-\pi}^{\pi} f^*(\theta)\, K(\theta)\, f(\theta)\, d\theta \tag{33}$$

$$= \sum_{r,k=0}^{m} \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-j(r-k)\theta}\, c_k^* K(\theta)\, c_r\, d\theta \tag{34}$$

$$= \sum_{r,k=0}^{m} c_k^* A_{r-k} c_r = c^* T_m c \tag{35}$$

where

$$c = \begin{bmatrix} c_0 \\ c_1 \\ . \\ . \\ . \\ c_m \end{bmatrix} \tag{36}$$

Hence, since c is arbitrary, T_m is nonnegative-definite, $m = 0 \to \infty$, Q.E.D.

Assume conversely that all T_m are nonnegative-definite and pick any constant n-vector x, any ρ, $0 < \rho < 1$, and any θ_o in I. Select the c_r so that

$$f(\theta) = x(1 - \rho^2)^{1/2} \sum_{r=0}^{m} \rho^r e^{-jr(\theta-\theta_o)} \quad , \tag{37}$$

i.e.,

$$c_r = x(1 - \rho^2)^{1/2} \rho^r e^{jr\theta_o} \quad , \quad r = 0 \to m \quad . \tag{38}$$

Now T_m nonnegative-definite implies that

$$0 \leq c^* T_m c = \frac{1}{2\pi} \int_{-\pi}^{\pi} f^*(\theta) K(\theta) f(\theta) \, d\theta = \frac{1}{2\pi} \int_{-\pi}^{\pi} (1 - \rho^2) \left| \sum_{r=0}^{m} \rho^r e^{-jr(\theta-\theta_o)} \right|^2 x^* K(\theta) x \, d\theta \quad . \tag{39}$$

Letting $m \to \infty$ in Eq. (39) and using dominated convergence,

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{1 - \rho^2}{1 - 2\rho \cos(\theta-\theta_o) + \rho^2} \cdot x^* K(\theta) x \, d\theta \geq 0 \quad . \tag{40}$$

The left-hand side of Eq. (40) is Poisson's integral and as is well known,[15] its limit as $\rho \to 1 - 0$ is equal to $x^* K(\theta_o) x$ for almost all θ_o. Thus, a.e.,

$$x^* K(\theta) x \geq 0 \quad . \tag{40a}$$

The exceptional set can depend on the choice of x. However, since a countable union of sets of measure zero is a set of measure zero, Eq. (40a) certainly holds a.e. for all rational x. But the latter are dense in the space of all n-vectors x and therefore $K(\theta) \geq 0$, a.e., Q.E.D.

Lastly, suppose some T_m is singular. Then there exists a nontrivial constant vector c of the form (36) such that $T_m c = 0$. Consequently, if

$$f(\theta) = \sum_{r=0}^{m} c_r e^{-jr\theta}$$

is constructed with this c, Eq. (33) to Eq. (35) show that

$$0 = \frac{1}{2\pi} \int_{-\pi}^{\pi} f^*(\theta) K(\theta) f(\theta) \, d\theta \quad ;$$

or, since $K(\theta) \geq 0$, a.e.,

$$K(\theta) f(\theta) = 0 \text{ a.e.}$$

But $f(\theta) \neq 0$ a.e. whence,

$$\det K(\theta) = 0$$

a.e. and

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \ln \det K(\theta) \, d\theta = -\infty \quad ,$$

contradicting Eq. (31). Assertion (2) now follows immediately and lemma 1 is established, Q.E.D.

The standard matrix Wiener-Hopf factorization problem can be phrased as follows. Determine the necessary and sufficient conditions for the existence of an n x n analytic matrix function

$$B(z) = \sum_{r=0}^{\infty} B_r z^r \tag{41}$$

of the class H_2 such that

$$\det B(z) \neq 0, \quad |z| < 1 \tag{42}$$

and

$$K(\theta) = B^*(e^{-j\theta}) \, B(e^{-j\theta}) \text{ a.e.} \tag{43}$$

Of course, the answer is well known.[13,I] Such a "minimum-phase" factor is unique up to multiplication on the left by an arbitrary constant n x n unitary matrix and exists iff

(a) $K(\theta)$ is hermitian nonnegative-definite, belongs to L_1 and satisfies the

the regularity condition*

$$(b)\quad \frac{1}{2\pi}\int_{-\pi}^{\pi} \ln \det K(\theta)\, d\theta > -\infty \quad . \tag{44}$$

The necessity of (a) is obvious and that of (b) is the subject of lemma 2.

Lemma 2.[13,I] If B(z) is any H_2-factor (41) satisfying Eqs. (42) and (43),

$$-\infty < \ln|\det B_0|^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} \ln \det K(\theta)\, d\theta \quad . \tag{45}$$

Proof. Clearly, det B(z) is free of zeros in $|z| < 1$ and

$$\det K(\theta) = |\det B(e^{-j\theta})|^2 \quad . \tag{46}$$

To reach Eq. (45) observe first that $\det B(0) = \det B_0 \neq 0$ and then apply the equality part of Jensen's inequality, Q.E.D.

Theorem 1. The n x n matrix coefficients B_r, $r = 0 \to \infty$, in the power series expansion (41) of the minimum-phase analytic n x n factor B(z) satisfying Eqs. (42) and (43) can be determined by means of the following two-step algorithm.

Step 1: For every $m \geq 0$ effect the (unique) Gauss factorization

$$T_m = L_m^* L_m \tag{47}$$

where

$$L_m = \begin{bmatrix} L_{00}^{(m)} & & & & \\ L_{10}^{(m)} & L_{11}^{(m)} & & & \\ \cdot & \cdot & \cdot & & \\ \cdot & \cdot & & \cdot & \\ \cdot & \cdot & & & \cdot \\ L_{m0}^{(m)} & L_{m1}^{(m)} & \cdot\ \cdot\ \cdot & & L_{mm}^{(m)} \end{bmatrix} \tag{48}$$

is square and lower-triangular with positive diagonal scalar entries. (All blocks in

---

*The inequality

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln \det K(\theta)\, d\theta < +\infty \tag{44a}$$

is automatic[13,II] and $K(\theta) \in L_1$ plus the regularity condition (44) imply $\ln \det K(\theta) \in L_1$.

Eq. (48) are n x n and all $L_{rr}^{(m)}$, $r = 0 \to m$, are themselves lower-triangular.)

Step 2: For every fixed r and k, $r \geq k \geq 0$,

$$\lim_{m\to\infty} L_{rk}^{(m)} = B_{r-k} \quad . \tag{49}$$

In particular,

$$\lim_{m\to\infty} L_{r0}^{(m)} = B_r, \quad r = 0 \to \infty \quad . \tag{50}$$

Furthermore, for any fixed $\ell \geq 0$, every diagonal scalar sequence $\{(L_{m\ell})_{i,i}\}$ of the matrix

$$L_{m\ell} = \begin{bmatrix} L_{00}^{(m)} & & & & \\ L_{10}^{(m)} & L_{11}^{(m)} & & & \\ . & . & . & & \\ . & . & & . & \\ . & . & & & . \\ L_{\ell 0}^{(m)} & L_{\ell 1}^{(m)} & . & . & . \; L_{\ell\ell}^{(m)} \end{bmatrix} \tag{51}$$

is monotonically nonincreasing while for $i \neq j$ and all A_r real, $r = 0 \to \infty$, $\{(L_{m\ell})_{i,j}^2\}$ is the difference of two monotonically nonincreasing sequences.

Proof. A 1-sided n x n trigonometric polynomial $P(\theta)$ of degree $\leq m$ has the form

$$P(\theta) = \sum_{r=0}^{m} X_r e^{-jr\theta} \tag{52}$$

where the X_r's are constant n x n matrices, $r = 0 \to m$, and

$$X = \begin{bmatrix} X_0 \\ X_1 \\ . \\ . \\ . \\ X_m \end{bmatrix} \tag{53}$$

is its associated n(m+1)xn coefficient matrix. Evidently, $P(\theta) = Q(e^{-j\theta})$ is the boundary value of

$$Q(z) = \sum_{r=0}^{m} X_r z^r \tag{54}$$

at $z = e^{-j\theta}$.

Suppose $X_0, X_1, \ldots, X_\ell$ are prescribed in advance, $\ell \leq m$, and put

$$X_a = \begin{bmatrix} X_0 \\ X_1 \\ . \\ . \\ . \\ X_\ell \end{bmatrix} \; ; \; X_b = \begin{bmatrix} X_{\ell+1} \\ X_{\ell+2} \\ . \\ . \\ . \\ X_m \end{bmatrix} \; ; \; X = \begin{bmatrix} X_a \\ \hline X_b \end{bmatrix} \quad . \tag{55}$$

For given m, ℓ and X_a, let $\mathcal{P}_{m,\ell}(X_a)$ denote the collection of all polynomials (52) whose coefficient matrices X have the specified upper part X_a. Thus, $\mathcal{P}_{m,\ell}(X_a)$ is generated by assigning the lower part of X arbitrarily and is therefore parametrized by X_b.

The problem of minimizing the quadratic functional

$$I(P) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathrm{Tr}\, \{P^*(\theta)\, K(\theta)\, P(\theta)\}\, d\theta \tag{56}$$

over P's belonging to $\mathcal{P}_{m,\ell}(X_a)$ is a basic one and its solution is fundamental to our entire approach. Let

$$\mu_{m,\ell}(X_a) = \underset{P \varepsilon \mathcal{P}_{m,\ell}(X_a)}{\text{minimum}} \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathrm{Tr}\, \{P^*(\theta)\, K(\theta)\, P(\theta)\}\, d\theta \quad . \tag{57}$$

Clearly, for $k \geq m$, $P \varepsilon \mathcal{P}_{m,\ell}(X_a)$ implies $P \varepsilon \mathcal{P}_{k,\ell}(X_a)$ and

$$\mu_{m,\ell}(X_a) \geq \mu_{k,\ell}(X_a) \geq 0 \quad . \tag{58}$$

Consequently, the sequence $\mu_{m,\ell}(X_a)$ is monotone nonincreasing with respect to m and the existence of

$$\mu_\ell(X_a) = \lim_{m \to \infty} \mu_{m,\ell}(X_a) \geq 0 \tag{59}$$

is assured for every fixed $\ell \geq 0$ and X_a.

To derive an explicit expression for $\mu_{m,\ell}(X_a)$ we first partition L_m in the conformable manner

$$L_m = \left[\begin{array}{cccc|c} L_{00}^{(m)} & & & & \\ L_{10}^{(m)} & L_{11}^{(m)} & & & \\ \vdots & \vdots & \ddots & & \\ L_{\ell 0}^{(m)} & L_{\ell 1}^{(m)} & \cdots & L_{\ell\ell}^{(m)} & \\ \hline & L_{ma} & & & L_{mb} \end{array}\right] = \left[\begin{array}{c|c} L_{m\ell} & \\ \hline L_{ma} & L_{mb} \end{array}\right] \tag{60}$$

and then verify the formula

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} P^*(\theta)\, K(\theta)\, P(\theta)\, d\theta = X^* T_m X \tag{61}$$

to obtain

$$I(P) = \mathrm{Tr}\,\{X^* T_m X\} = \mathrm{Tr}\,\{(L_m X)^* (L_m X)\} \quad . \tag{62}$$

Since

$$L_m X = \left[\begin{array}{c} L_{m\ell} X_a \\ \hline L_{ma} X_a + L_{mb} X_b \end{array}\right] \quad , \tag{63}$$

$$I(P) = \mathrm{Tr}\,\{X_a^*(L_{m\ell}^* L_{m\ell})X_a + (L_{ma}X_a + L_{mb}X_b)^*(L_{ma}X_a + L_{mb}X_b)\}$$

$$\geq \mathrm{Tr}\,\{X_a^*(L_{m\ell}^* L_{m\ell})X_a\} \quad . \tag{64}$$

Thus

$$\mu_{m,\ell}(X_a) = \mathrm{Tr}\,\{X_a^* \Gamma_{m,\ell} X_a\} \tag{65}$$

and

$$\mu_\ell(X_a) = \lim_{m\to\infty} \mathrm{Tr}\,\{X_a^* \Gamma_{m,\ell} X_a\} \quad , \tag{66}$$

$$\Gamma_{m,\ell} = L_{m\ell}^* L_{m\ell} \quad . \tag{67}$$

The minimum $\mu_{m,\ell}(X_a)$ is achieved iff*

$$X_b = -L_{mb}^{-1} L_{ma} X_a \quad . \tag{68}$$

---

* $0 \neq \sqrt[+]{\det T_m} = \det L_m = (\det L_{m\ell}) \cdot (\det L_{mb})$. (68a)

Let X_{a1} and X_{a2} be any two fixed choices of X_a. From the convergence of the three sequences

$$\{Tr(X_{a1}^* \Gamma_{m,\ell} X_{a1})\},\ \{Tr(X_{a2}^* \Gamma_{m,\ell} X_{a2})\},\ \{Tr((X_{a1}+X_{a2})_* \Gamma_{m,\ell} (X_{a1}+X_{a2}))\}$$

we infer that of $\{\text{Re } Tr(X_{a1}^* \Gamma_{m,\ell} X_{a2})\}$ as the difference of two monotone nonincreasing sequences. Similarly, replacing X_{a1} by jX_{a1} we infer that of $\{\text{Im Re}(X_{a1}^* \Gamma_{m,\ell} X_{a2})\}$ and finally that of $\{Tr(X_{a1}^* \Gamma_{m,\ell} X_{a2})\}$ as the difference of two (possibly complex) sequences whose respective real and imaginary parts of monotonically nonincreasing. In the special but important case that all A_r are real, $r = 0 \to \infty$, all matrices $\Gamma_{m,\ell}$ are real and $\{Tr(X_{a1}^* \Gamma_{m,\ell} X_{a2})\}$ is the difference of two monotone nonincreasing sequences for every real pair (X_{a1}, X_{a2}).

By a suitable choice of X_{a1} and X_{a2}, $Tr(X_{a1}^* \Gamma_{m,\ell} X_{a2})$ can be made equal to any scalar entry in $\Gamma_{m,\ell}$ and therefore

$$\Gamma_\ell = \lim_{m\to\infty} \Gamma_{m,\ell} \tag{69}$$

exists for every fixed $\ell \geq 0$. More strongly, every diagonal scalar element in $\Gamma_{m,\ell}$ approaches its limit in a monotone nonincreasing manner and if all A_r are real, every off-diagonal element converges as the limit of the difference of two monotone nonincreasing sequences.

Although it is evident that the limit matrix Γ_ℓ is hermitian nonnegative definite, it is not at all obvious that it is actually positive-definite. From the formula

$$\sqrt[+]{\det \Gamma_{m,\ell}} = \prod_{r=0}^{\ell} \det L_{rr}^{(m)} \quad , \tag{70}$$

it is seen that $\det \Gamma_\ell = 0$ implies

$$\inf_m \det L_{rr}^{(m)} = 0 \tag{71}$$

for at least one r, $0 \leq r \leq \ell$. However, by examining the lower right-hand corner $n(m+1)\times n(m+1)$ block of T_{m+1} (use Eqs. (47) and (48)), it is easily seen that

$$L_{00}^{(m)} = L_{11}^{(m+1)} \quad \text{and} \quad L_{10}^{(m)} = L_{21}^{(m+1)} \quad . \tag{72}$$

In general, for given $q \geq 0$ and $r \geq k$,

$$L_{rk}^{(m)} = L_{r+q,k+q}^{(m+q)} \tag{73}$$

for m sufficiently large. Thus, if Eq. (71) is satisfied for one value of r it is satisfied

for every $r = 0 \to \ell$ and $\det \Gamma_\ell = 0$ implies

$$\inf_m \det L_{00}^{(m)} = 0 \tag{74}$$

which, as we will show, is impossible.

Consider the problem of minimizing I(P) over $\mathcal{P}_{m,0}(1_n)$ and let $P_m(\theta)$ denote the corresponding extremalizing polynomial. Note that $X_a = X_0 = 1_n$ and $P_m(\theta) = Q_m(e^{-j\theta})$ where

$$Q_m(z) = 1_n + \sum_{r=1}^{m} X_r z^r \quad . \tag{75}$$

From Eqs. (64) and (67),

$$(L_{00}^{(m)})^* L_{00}^{(m)} = \Gamma_{m,0} = \frac{1}{2\pi} \int_{-\pi}^{\pi} P_m^*(\theta)\, K(\theta)\, P_m(\theta)\, d\theta \tag{76}$$

and applying the master inequality and Jensen's inequality (section A) in succession,

$$\ln \det \Gamma_{m,0} \geq \frac{1}{2\pi} \int_{-\pi}^{\pi} \ln \det K(\theta)\, d\theta + \frac{1}{2\pi} \int_{-\pi}^{\pi} \ln \left| \det P_m(\theta) \right|^2 d\theta \geq \frac{1}{2\pi} \int_{-\pi}^{\pi} \ln \det K(\theta)\, d\theta \tag{77}$$

since $\ln \det Q_m(0) = \ln \det 1_n = 0$. Thus, imposing the regularity condition (44),

$$\inf_m \det \Gamma_{m,0} \geq \exp\left[\frac{1}{2\pi} \int_{-\pi}^{\pi} \ln \det K(\theta)\, d\theta \right] > 0 \quad . \tag{78}$$

$$\therefore \quad \inf_m \det L_{00}^{(m)} = \inf_m \sqrt{\det \Gamma_{m,0}} > 0 \tag{79}$$

and assumption (74) is untenable, Q.E.D.

The hermitian positive-definite character of the limit matrix Γ_ℓ now permits an easy proof of the existence of

$$L_\ell = \lim_{m \to \infty} L_{m\ell} \quad . \tag{80}$$

Let

$$\Gamma_\ell = \Delta_\ell^* \Delta_\ell \tag{81}$$

where Δ_ℓ is the unique square, lower-triangular factor with positive diagonal elements and define

$$U_{m\ell} = L_{m\ell}\,\Delta_{\ell}^{-1} \tag{82}$$

Observe that $U_{m\ell}$ is also square and lower-triangular with positive diagonal elements. From Eq. (67),

$$\underset{m\to\infty}{\text{limit}}(U^{*}_{m\ell}\,U_{m\ell}) = 1 \tag{83}$$

and a straightforward argument taking the properties of $U_{m\ell}$ into account yields

$$\underset{m\to\infty}{\text{limit}}\; U_{m\ell} = 1 \quad . \tag{84}$$

Or, from Eq. (82),

$$\underset{m\to\infty}{\text{limit}}\; L_{m\ell} = \Delta_{\ell} \quad . \tag{85}$$

By selecting ℓ large enough any block $L^{(m)}_{rk}$, $r \geq k$, can be encompassed within $L_{m\ell}$ and therefore

$$\underset{m\to\infty}{\text{limit}}\; L^{(m)}_{rk} = B_{rk} \tag{86}$$

exists, Q.E.D.

In view of Eq. (73),

$$\underset{m\to\infty}{\text{limit}}\; L^{(m+q)}_{r+q,\,k+q} = B_{r+q,\,k+q} = \underset{m\to\infty}{\text{limit}}\; L^{(m)}_{rk} = B_{rk}$$

whence,

$$B_{rk} = \underset{m\to\infty}{\text{limit}}\; L^{(m)}_{rk} = B_{r-k},\quad r \geq k \quad . \tag{87}$$

In particular,

$$\underset{m\to\infty}{\text{limit}}\; L^{(m)}_{r0} = B_{r},\quad r = 0 \to \infty \quad . \tag{88}$$

Thus, for any fixed $\ell \geq 0$,

$$\Delta_{\ell} = \underset{m\to\infty}{\text{limit}}\; L_{m\ell} = \begin{bmatrix} B_0 & & & & \\ B_1 & B_0 & & & \\ \cdot & \cdot & \cdot & & \\ \cdot & \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ B_{\ell} & B_{\ell-1} & \cdot & \cdot & B_0 \end{bmatrix} \quad . \tag{89}$$

By appealing to the already established convergence properties of $\{\Gamma_{m\ell}\}$ and a slight generalization of the minimum problem* we can readily demonstrate that every diagonal scalar sequence

$$\{(L_{m\ell})_{i,i}\}$$

is monotonically nonincreasing while for all $i \neq j$ and all A_r real, $r = 0 \to \infty$,

$$\{(L_{m\ell})^2_{i,j}\}$$

is the difference of two monotonically nonincreasing sequences. By going to the limit in Eq. (77) we derive the important inequality (actually an equality),

$$\ln|\det B_0|^2 \geq \frac{1}{2\pi}\int_{-\pi}^{\pi} \ln \det K(\theta)\, d\theta \quad . \tag{90}$$

From Eqs. (47) and (48),

$$A_0 = \sum_{r=0}^{m} (L_{r0}^{(m)})^* L_{r0}^{(m)} \tag{90a}$$

which implies that

$$A_0 - \sum_{r=0}^{\ell} (L_{r0}^{(m)})^* L_{r0}^{(m)} \geq 0 \tag{91}$$

for every fixed $\ell \leq m$. Letting $m \to \infty$,

$$A_0 - \sum_{r=0}^{\ell} B_r^* B_r \geq 0 \tag{92}$$

which assures both the existence of the infinite sum

$$\sum_{r=0}^{\infty} B_r^* B_r$$

and the inequality

$$A_0 - \sum_{r=0}^{\infty} B_r^* B_r \geq 0 \quad . \tag{93}$$

---

*Instead of specifying the $n \times n$ blocks $X_0, X_1, \ldots, X_\ell$ in advance we fix the first $\ell+1$ scalar rows (from the top) of the coefficient matrix X.

The matrix function

$$B(z) = \sum_{r=0}^{\infty} B_r z^r \tag{94}$$

is therefore analytic and of class H_2 in $|z| < 1$.

Generalizing, from $T_m = L_m^* L_m$ and Eq. (60) we obtain

$$T_\ell = L_{m\ell}^* L_{m\ell} + L_{ma}^* L_{ma} \tag{95}$$

which, when subjected to the same limiting process as above gives

$$T_\ell - (B^{(\ell)})^* B^{(\ell)} \geq 0 \quad . \tag{96}$$

Here

$$B^{(\ell)} = \begin{bmatrix} B_0 & & & & & \\ B_1 & B_0 & & & & \\ \cdot & \cdot & \cdot & & & \\ \cdot & \cdot & \cdot & \cdot & & \\ \cdot & \cdot & \cdot & & \cdot & \\ B_\ell & B_{\ell-1} & \cdot & \cdot & \cdot & B_0 \\ \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot \end{bmatrix} \tag{97}$$

has $\ell+1$ block columns and an infinite number of block rows. But

$$(B^{(\ell)})^* B^{(\ell)}$$

is nothing more than the corresponding Toeplitz matrix (30) generated by the Fourier coefficients of $B^*(e^{-j\theta}) B(e^{-j\theta})$ and invoking lemma 1, Eq. (96) yields

$$K(\theta) - B^*(e^{-j\theta}) B(e^{-j\theta}) \geq 0 \quad \text{a.e.} \tag{98}$$

Thus, from Eq. (19),

$$\ln \det K(\theta) \geq \ln |\det B(e^{-j\theta})|^2 \tag{99}$$

a.e. and Jensen's inequality plus Eq. (90) give

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \ln \det K(\theta)\, d\theta \geq \frac{1}{2\pi} \int_{-\pi}^{\pi} \ln |\det B(e^{-j\theta})^2 d\theta \geq \ln |\det B_0|^2 \geq \frac{1}{2\pi} \int_{-\pi}^{\pi} \ln \det K(\theta)\, d\theta \quad . \tag{100}$$

Consequently, all four terms in Eq. (100) are equal. From the equality of the inner two we deduce (Jensen) that det B(z) is free of zeros in $|z| < 1$ and from the equality of the two on the left and Eq. (99),

$$\det K(\theta) = |\det B(e^{-j\theta})|^2 \quad , \quad \text{a.e.} \tag{101}$$

However, because of Eq. (98), the a.e. nonsingularity of $K(\theta)$ and the hermitian non-negative-definite character of both $K(\theta)$ and $B^*(e^{-j\theta})B(e^{-j\theta})$, Eq. (101) is possible iff

$$K(\theta) = B^*(e^{-j\theta})B(e^{-j\theta}) \quad \text{a.e.} \tag{102}$$

The proof of theorem 1 is complete, Q.E.D.

C. Matrix Polynomials Orthogonal on the Unit Circle

Before introducing these polynomials we shall establish an important property of the $Q_m(z)$. Recall, that $P_m(\theta)$, the unique n x n trigonometric polynomial minimizing I(P) over $\mathcal{P}_{m,0}(1_n)$ equals $Q_m(e^{-j\theta})$ where

$$Q_m(z) = 1_n + \sum_{r=1}^{m} X_r z^r \quad ,$$

$m = 0 \rightarrow \infty$ and the X_r are chosen in accordance with Eq. (68). It is convenient to refer to $Q_m(z)$ as the mth generating polynomial over $\mathcal{P}_{m,0}(1_n)$.

Lemma 3. The matrix polynomial $Q_m(z)$ is minimum-phase; i.e.,

$$\det Q_m(z) \neq 0, \quad |z| < 1 \quad , \tag{103}$$

$m = 0 \rightarrow \infty$.

Proof. Suppose that $\det Q_m(z_o) = 0$, $|z_o| < 1$. Then, since $Q_m(0) = 1_n$, $z_o \neq 0$ and for some n-vector x satisfying $x^*x = 1$,

$$Q_m(z_o)x = 0 \quad . \tag{104}$$

An easy check shows that the matrix

$$\Lambda(z) = 1_n - \frac{1 - |z_o|^2}{1 + z_o} \cdot \frac{1 + z}{1 - zz_o^*} xx^* \tag{105}$$

is regular paraconjugate unitary; i.e., $\Lambda(z)$ is analytic in $|z| \leq 1$ and

$$\Lambda(z)\,\Lambda_*(z) = 1_n \quad . \tag{106}$$

(For $z = e^{-j\theta}$, $\Lambda(e^{-j\theta})$ is unitary.) Noting that

$$\det \Lambda(0) = z_o \cdot \frac{1 + z_o^*}{1 + z_o} \neq 0 \quad , \tag{107}$$

it is obvious that

$$\Lambda_N(z) = \Lambda^{-1}(0)\,\Lambda(z) \tag{108}$$

is analytic in $|z| \leq 1$ and normalized to 1_n at $z = 0$.

Let

$$Q(z) = Q_m(z)\,\Lambda_N^{-1}(z)$$

$$= Q_m(z)\left(1_n - \frac{1 - |z_o|^2}{1 + z_o^*} \cdot \frac{1 + z}{z - z_o} xx^*\right)\Lambda(0) \quad . \tag{109}$$

Clearly, $Q(0) = 1_n$ and the residue matrix at the only possible finite pole $z = z_o$ is proportional to $Q_m(z_o)xx^* = 0$. Thus $Q(z)$ is an $n \times n$ polynomial matrix in z of degree $\leq m$ and $P(\theta) = Q(e^{-j\theta}) \in \mathcal{P}_{m,0}(1_n)$. But the correlated cost

$$I(P) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{Tr}\,\{P^*(\theta)\,K(\theta)\,P(\theta)\}\,d\theta$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{Tr}\,\{\Lambda^*(0)\,\Lambda(e^{-j\theta})P_m^*(\theta)\,K(\theta)\,P_m(\theta)\,\Lambda^*(e^{-j\theta})\,\Lambda(0)\}\,d\theta$$

$$\leq \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{Tr}\,\{\Lambda(e^{-j\theta})P_m^*(\theta)\,K(\theta)\,P_m(\theta)\,\Lambda^*(e^{-j\theta})\}\,d\theta$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{Tr}\,\{P_m^*(\theta)\,K(\theta)\,P_m(\theta)\}\,d\theta$$

$$= I(P_m) \leq I(P) \quad .$$

(In going from line 2 to the inequality in line 3 we have made use of the fact that $\Lambda(0)\,\Lambda^*(0)$ possesses only the two distinct eigenvalues 1 and $|z_o|^2 < 1$.) Hence, $I(P) = I(P_m)$ and by uniqueness, $P(\theta) = P_m(\theta)$ which, in view of Eq. (109) is absurd unless x is the zero vector, a contradiction, Q.E.D.

Evidently, because of the extremalizing property of $P_m(\theta)$, for every $m \geq 0$

$$P(\theta) = e^{jm\theta}\,P_m(\theta) \tag{110}$$

is the unique trigonometric polynomial of the form

$$1_n \cdot e^{jm\theta} + \sum_{r=0}^{m-1} X_r e^{j(m-r)\theta} \tag{111}$$

minimizing I(P) and is realized as the boundary value at $z = e^{j\theta}$ of the monic polynomial

$$Q^{(m)}(z) \equiv z^m Q_m(\tfrac{1}{z}) = z^m \cdot 1_n + \sum_{r=0}^{m-1} z^r Q_r^{(m)} \quad , \tag{112}$$

$$Q_r^{(m)} = X_{m-r} \ , \ r = 0 \to m-1 \quad . \tag{113}$$

More important is the observation that the sequence $\{Q^{(m)}(z)\}$ is matrix left-orthogonal* for $K(\theta)$; i.e.,

$$\langle Q^{(r)}(z), Q^{(k)}(z) \rangle \equiv \frac{1}{2\pi} \int_{-\pi}^{\pi} Q_*^{(r)}(e^{j\theta}) K(\theta) Q^{(k)}(e^{j\theta}) \, d\theta = O_n \quad , \tag{114}$$

$r \neq k = 0 \to \infty$.

Proof. From the identification (112) and Eqs. (63) + (68) we obtain the formula**

$$L_m \Omega_m = L_{00}^{(m)} \dot{+} L_{00}^{(m-1)} \dot{+} \ldots \dot{+} L_{00}^{(0)} \tag{115}$$

where

$$\Omega_m = \begin{bmatrix} 1_n & & & & \\ Q_{m-1}^{(m)} & 1_n & & & \\ Q_{m-2}^{(m)} & Q_{m-2}^{(m-1)} & & & \\ \cdot & \cdot & \cdot & & \\ \cdot & \cdot & & \cdot & \\ \cdot & \cdot & & & \cdot \\ Q_0^{(m)} & Q_0^{(m-1)} & \cdot & \cdot & \cdot \quad 1_n \end{bmatrix} \ , \tag{116}$$

$m = 0 \to \infty$. Thus, all the off-diagonal n x n blocks of

---

*$K(\theta)$ is assumed to possess all the properties enumerated in theorem 1.

**$A \dot{+} B$ is the "direct" sum of matrices A and B.

$$\Omega_m^*(L_m^* L_m)\Omega_m = \Omega_m^* T_m \Omega_m \tag{117}$$

equal O_n which is precisely what is required for the validity of Eq. (114), Q.E.D.

According to Eq. (115),

$$\begin{bmatrix} I_n \\ Q_{m-1}^{(m)} \\ Q_{m-2}^{(m)} \\ \vdots \\ Q_0^{(m)} \end{bmatrix} = (L_m^{-1})_{;1} \cdot L_{00}^{(m)} \quad , \tag{118}$$

$(L_m^{-1})_{;1}$ denoting the first block column of the inverse of L_m. Hence, Eq. (118) can be used to generate the coefficients of all the $Q^{(m)}(z)$, $m = 0 \to \infty$.

In the same way, the n x n matrix polynomials

$$R^{(m)}(z) = z^m \cdot I_n + \sum_{r=0}^{m-1} z^r R_r^{(m)} \quad , \tag{119}$$

$m = 0 \to \infty$, are said to be right-orthogonal for $K(\theta)$ if

$$\langle R^{(r)}(z), R^{(k)}(z) \rangle \equiv \frac{1}{2\pi} \int_{-\pi}^{\pi} R^{(r)}(e^{j\theta}) K(\theta) R_*^{(k)}(e^{j\theta}) \, d\theta = O_n \quad , \tag{120}$$

$r \neq k = 0 \to \infty$.

It follows immediately that the sequence $\{R^{(m)'}(z)\}$ is left-orthogonal for $K'(\theta)$ and in particular, if $K(\theta)$ is symmetric,

$$R^{(m)}(z) = Q^{(m)'}(z), \quad m = 0 \to \infty \quad . \tag{121}$$

In general if

$$\tilde{T}_m \equiv \begin{bmatrix} A_0 & A_{-1} & \cdots & A_{-m} \\ A_1 & A_0 & \cdots & A_{-m+1} \\ \vdots & \vdots & & \vdots \\ A_m & A_{m-1} & \cdots & A_0 \end{bmatrix} = M_m^* M_m \tag{122}$$

with

$$M_m = \begin{bmatrix} M_{00}^{(m)} & & & & \\ M_{10}^{(m)} & M_{11}^{(m)} & & & \\ \cdot & \cdot & \cdot & & \\ \cdot & \cdot & & \cdot & \\ \cdot & \cdot & & & \cdot \\ M_{m0}^{(m)} & M_{m1}^{(m)} & \cdot & \cdot & \cdot \; M_{mm}^{(m)} \end{bmatrix} \tag{123}$$

chosen as the unique lower-triangular factor with positive diagonal scalar elements,

$$\begin{bmatrix} 1_n \\ R_{m-1}^{(m)*} \\ R_{m-2}^{(m)*} \\ \cdot \\ \cdot \\ \cdot \\ R_0^{(m)*} \end{bmatrix} = (M_m^{-1})_{;1} \cdot M_{00}^{(m)} , \quad m = 0 \to \infty \quad . \tag{124}$$

When all A_r's are real and symmetric (which is true if $K(\theta)$ is real and even), the concepts of left and right-orthogonal on the unit circle simplify and

$$R^{(m)}(z) = Q^{(m)'}(z) , \quad m = 0 \to \infty \quad . \tag{125}$$

Equating the corresponding $(\ell+1)n \times (\ell+1)n$ upper left-hand corner blocks on both sides of Eq. (115) yields

$$L_{m\ell} \cdot \Omega_{m\ell} = L_{00}^{(m)} \dot{+} L_{00}^{(m-1)} \dot{+} \ldots \dot{+} L_{00}^{(m-\ell)} \tag{126}$$

where

$$\Omega_{m\ell} = \begin{bmatrix} 1_n & & & & \\ Q_{m-1}^{(m)} & 1_n & & & \\ Q_{m-2}^{(m)} & Q_{m-2}^{(m-1)} & & & \\ \cdot & \cdot & \cdot & & \\ \cdot & \cdot & & \cdot & \\ \cdot & \cdot & & & \cdot \\ Q_{m-\ell}^{(m)} & Q_{m-\ell}^{(m-1)} & \cdot & \cdot & \cdot \; 1_n \end{bmatrix} . \tag{127}$$

Let $m \to \infty$ in Eq. (126) with ℓ held fixed and define

$$\Omega_\ell = \lim_{m\to\infty} \Omega_{m\ell} \quad . \tag{128}$$

Then, using Eq. (89),

$$\Delta_\ell \cdot \Omega_\ell = B_0 \dotplus B_0 \dotplus \ldots \dotplus B_0 \tag{129}$$

and we have derived the remarkable result

$$\Omega_\ell = \begin{bmatrix} 1_n & & & & \\ B_0^{-1}B_1 & 1_n & & & \\ B_0^{-1}B_2 & B_0^{-1}B_1 & & & \\ \cdot & \cdot & \cdot & & \\ \cdot & \cdot & & \cdot & \\ \cdot & \cdot & & & \cdot \\ B_0^{-1}B_\ell & B_0^{-1}B_{\ell-1} & \cdot\ \cdot\ \cdot & & 1_n \end{bmatrix}^{-1} \quad . \tag{130}$$

Equivalently, for $\ell = 0 \to \infty$ all limits

$$C_\ell = \lim_{m\to\infty} Q^{(m)}_{m-\ell} \tag{131}$$

exist and are given explicitly by the corresponding entries in the first block column of the inverse matrix appearing in Equation (130). Furthermore, as should be clear from Eq. (129), the power series

$$C(z) = 1_n + \sum_{r=1}^{\infty} C_r z^r \tag{132}$$

provides the inverse of the normalized canonic factor $B_0^{-1}B(z)$:

$$(1_n + B_0^{-1}B_1 z + B_0^{-1}B_2 z^2 + \ldots + ..)(1_n + C_1 z + C_2 z^2 + \ldots + ..) = 1_n \quad . \tag{133}$$

Thus, the limiting values (131) of the coefficients of the left-orthogonal polynomials $Q^{(m)}(z)$ can be employed to construct the normalized canonic factor $B_0^{-1}B(z)$ or vice-versa! For example,

$$\begin{aligned} -B_0^{-1}B_1 &= C_1 \quad , \\ -B_0^{-1}B_2 &= C_2 - C_1^2 \quad , \end{aligned} \tag{134}$$

etc., involving the inversion of only one n x n matrix

$$B_0 = \lim_{m\to\infty} L_{00}^{(m)} \quad . \tag{134a}$$

D. Recursion Formulas for the Orthogonal Polynomials

As is well known,[17] in the scalar case (n = 1) the $Q^{(m)}(z)$ are tied together by a recursion of the form

$$Q^{(m+1)}(z) = zQ^{(m)}(z) - z^m Q_*^{(m)}(z)\, a_m \quad , \tag{135}$$

$m = 0 \to \infty$. However, for $n > 1$ the situation is complicated by the existence of distinct left and right-orthogonal sets.

Theorem 2. Let $K(\theta)$ satisfy the conditions of theorem 1. Then,

(1) the left and right-orthogonal n x n matrix polynomials associated with $K(\theta)$ are generated by the recurrences

$$Q^{(m+1)}(z) = zQ^{(m)}(z) - z^m \cdot R_*^{(m)}(z) \cdot a_m \quad , \tag{136}$$

$$R^{(m+1)}(z) = zR^{(m)}(z) - z^m \cdot \beta_m \cdot Q_*^{(m)}(z) \quad , \tag{137}$$

where $Q^{(0)}(z) = R^{(0)}(z) = 1_n$ and

$$a_m = (M_{00}^{(m)*} M_{00}^{(m)})^{-1} \cdot (R^{(m)}K)_{-1} = (M_{00}^{(m)*} M_{00}^{(m)})^{-1} \cdot (KQ^{(m)})_{-1} \quad , \tag{138}$$

$$\beta_m = (R^{(m)}K)_{-1} \cdot (L_{00}^{(m)*} L_{00}^{(m)})^{-1} = (KQ^{(m)})_{-1} \cdot (L_{00}^{(m)*} L_{00}^{(m)})^{-1} \quad , \tag{139}$$

$m = 0 \to \infty$.

In Eqs. (138) and (139), $(R^{(m)}K)_k$ and $(KQ^{(m)})_k$ denote the constant n x n matrix coefficients of $e^{jk\theta}$ in the Fourier expansions of $R^m(e^{j\theta})K(\theta)$ and $K(\theta)Q^{(m)}(e^{j\theta})$, respectively. Hence, if m = 0

$$(R^{(0)}K)_{-1} = (K)_{-1} = A_1^* = (K)_{-1} = (KQ^{(0)})_{-1} \quad , \tag{140}$$

$$M_{00}^{(0)*} M_{00}^{(0)} = A_0 = L_{00}^{(0)*} L_{00}^{(0)} \quad , \tag{141}$$

$$a_0 = A_0^{-1}A_1^* \; ; \; \beta_0 = A_1^* A_0^{-1} \quad . \tag{142}$$

(2) For each $m = 0 \to \infty$, $Q^{(m)}(e^{j\theta})$ and $R^{(m)}(e^{j\theta})$ are the unique n x n trigonometric polynomials of the form

$$P(\theta) = 1_n \cdot e^{jm\theta} + \sum_{r=1}^{m} X_r e^{j(m-r)\theta} \tag{143}$$

minimizing the respective quadratic functionals

$$I_Q(P) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathrm{Tr}\,\{P^*(\theta)\, K(\theta)\, P(\theta)\}\, d\theta \tag{144}$$

and

$$I_R(P) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathrm{Tr}\,\{P(\theta)\, K(\theta)\, P^*(\theta)\}\, d\theta \quad . \tag{145}$$

(3) For $m = 0 \to \infty$,

$$\det Q^{(m)}(z) \cdot \det R^{(m)}(z) \neq 0 \ , \quad |z| > 1 \quad . \tag{146}$$

Proof. Only part (1) requires verification. Since

$$z^m 1_n - Q^{(m)}(z) \quad \text{and} \quad z^m 1_n - R^{(m)}(z)$$

are both of degree $\leq m-1$, it is clear that

$$\frac{z^m 1_n - Q^{(m)}(z)}{z^{m-1}} = \sum_{r=0}^{m-1} R_*^{(r)}(z)\, a_r^{(m)} \tag{147}$$

and

$$\frac{z^m 1_n - R^{(m)}(z)}{z^{m-1}} = \sum_{r=0}^{m-1} \beta_r^{(m)} Q_*^{(r)}(z) \tag{148}$$

for some choice of $n \times n$ constant matrices $a_r^{(m)}$, $\beta_r^{(m)}$, $r = 0 \to m-1$, $m = 1 \to \infty$. Now from the very definition of the orthogonal polynomials it follows that

$$(KQ^{(m)})_k = (R^{(m)}K)_k = O_n \ , \quad k = 0 \to m-1 \ , \tag{149}$$

$$(KQ^{(m)})_m = L_{00}^{(m)*} L_{00}^{(m)} \ ; \ (R^{(m)}K)_m = M_{00}^{(m)*} M_{00}^{(m)} \quad . \tag{150}$$

Thus

$$(R^{(r)}KQ^{(m)})_{m-1} = (R^{(m)}KQ^{(r)})_{m-1} = O_n \ , \ r = 0 \to m-1 \quad . \tag{151}$$

Setting $z = e^{j\theta}$, multiplying both sides of Eq. (147) on the left by $R^{(r)}(e^{j\theta})K(\theta)$ and then using the orthogonality relations (120) plus (150) and (151) gives

$$a_r^{(m)} = (M_{00}^{(r)*} M_{00}^{(r)})^{-1} \cdot (R^{(r)}K)_{-1} \tag{152}$$

Similarly, from Eq. (148),

$$\beta_r^{(m)} = (KQ^{(r)})_{-1} \cdot (L_{00}^{(r)*} L_{00}^{(r)})^{-1} \tag{153}$$

and both $a_r^{(m)}$ and $\beta_r^{(m)}$ are independent of m, $r = 0 \to \infty$; i.e., $a_r^{(m)} = a_r$ and $\beta_r^{(m)} = \beta_r$.

Equations (147) and (148) now read

$$Q^{(m)}(z) = z^m 1_n - z^{m-1} \cdot \sum_{r=0}^{m-1} R_*^{(r)}(z)\, a_r \quad , \tag{154}$$

$$R^{(m)}(z) = z^m 1_n - z^{m-1} \cdot \sum_{r=0}^{m-1} \beta_r Q_*^{(r)}(z) \tag{155}$$

and Eqs. (136), (137) drop out immediately.

Lastly, multiplying both sides of Eq. (137) on the right by $K(\theta)$ and integrating over $I = [-\pi \leq \theta \leq \pi]$ we get

$$\beta_m = (R^{(m)}K)_{-1} \cdot (L_{00}^{(m)*} L_{00}^{(m)})^{-1} \quad , \tag{156}$$

establishing Eqs. (138) and (139) Q. E. D.

National Science Foundation
ENG 74-13054

D. C. Youla and N. N. Kazanjian

REFERENCES

1. D.C. Youla, "On the Factorization of Rational Matrices," IRE Trans. Information Theory, IT-7, No. 3, 172-189 (July 1961).
2. E. Wong and J.B. Thomas, "On the Multidimensional Prediction and Filtering Problem and the Factorization of Spectral Matrices," J. Franklin Inst., 272, No. 2, 87-99 (August 1961).
3. A.C. Riddle and B.D. Anderson, "Spectral Factorization - Computational Aspects," IEEE Trans. Automatic Control, AC-11, 764-765 (October 1966).
4. F. Csáki and P. Fischer, "On the Spectrum Factorization," Acta Tech. (Budapest), 58, 145-168 (1967).
5. M.G. Strintzis, "A Solution to the Matrix Factorization Problem," IEEE Trans. Information Theory, IT-18, 225-232 (March 1972).
6. B.D.O. Anderson, "An Algebraic Solution to the Spectral Factorization Problem," IEEE Trans. Automatic Control, AC-12, 410-414 (August 1967).
7. W.G. Tuel, "Computer Algorithm for Spectral Factorization of Rational Matrices," IBM J. Res. Develop., 163-170 (March 1968).
8. G. Tunnicliffe Wilson, "The Factorization of Matricial Spectral Densities," SIAM J. Appl. Math., 23, No. 4, 420-426 (December 1972).

9. F. L. Bauer, "Ein direktes Iterationsverfahren zur Hurwitz-Zerlegung eines Polynoms," Archiv der Elektrischen Ubertragung, 9, 285-290 (1955).

10. J. Rissanen and T. Kailath, "Partial Realization of Random Systems," in 5th IFAC World Congress, (paper 35.6), Paris (1972).

11. J. von Neumann, "The Geometry of Orthogonal Spaces," II (Princeton University Press, 1950).

12. L. Mirsky, "Linear Algebra," (Clarendon Press, 1963).

13. N. Wiener and P. Masani, "The Prediction Theory of Multivariate Processes," Part I: "The Regularity Condition," Acta Math., 98 (1957); Part II: "The Linear Predictor," Acta Math., 99 (1958).

14. G.H. Hardy, J.E. Littlewood and G. Polya, "Inequalities," (Cambridge University Press, 1952).

15. R.P. Boas, "Entire Functions," (New York: Academic Press, 1954).

16. A. Zygmund, "Trigonometric Series," 2nd Ed., (Cambridge University Press, 1959).

17. U. Grenander and G. Szegö, "Toeplitz Forms and Their Applications," (University of California Press, 1958).

ON THE DESIGN OF SINGLE-LOOP SINGLE-INPUT-OUTPUT FEEDBACK CONTROL SYSTEMS IN THE COMPLEX-FREQUENCY DOMAIN

J.J. Bongiorno, Jr. and D.C. Youla

This summary is concerned with the single-loop, single-input-output, feedback control system shown in Figure 1. The subsystems are modeled by rational transfer functions and the transfer functions F(s), $F_o(s)$, P(s) and $P_o(s)$ are given. Attention is restricted to the large class of practical cases for which P(s) is proper* and $F(s)P(s)C(s) \not\equiv 0$.** The objective is the determination of the controller transfer function C(s) for which the system is asymptotically stable and design objectives are met. This has always been the objective in feedback system design. What is novel about the approach reported on here is that the family of all stabilizing controller transfer functions for the given F(s) and P(s) is identified and utilized. Specifically, every controller which stabilizes the loop is of the form[2]

$$C(s) = \frac{Y(s) + A(s)K(s)}{X(s) - B(s)K(s)} \quad , \tag{1}$$

where the polynomials A(s), B(s), X(s) and Y(s) are determined from the product F(s)P(s) -- the given data -- and K(s) is <u>any</u>*** rational function free of finite poles (analytic) in Re $s \geq 0$. Moreover, all the freedom that exists in selecting the closed-loop characteristic polynomial resides in the choice of denominator for K(s). The selection of this denominator is the first step in the design method described here and is done so that asymptotic stability as well as acceptable modal responses are assured. Hence, the struggle over stability everpresent in all classical design schemes is circumvented! The remaining design freedom lies in the selection of the numerator for K(s).

All the available design freedom (i.e., the selection of a rational K(s) analytic in Re $s \geq 0$) is utilized in Refs. 1 and 2 to obtain controllers which are optimal with respect to a performance index selected on the basis of sound engineering considerations. The optimal approach, however, requires that the system input signals be characterized by appropriate power spectral densities. These data are often not available or too costly to obtain. This justifies the attention which has been directed toward the selection of a suboptimal controller.

---

*A transfer function T(s) is proper if $T(\infty)$ is finite and strictly proper if $T(\infty) = 0$. Otherwise, it is improper.

**If $F(s)P(s)C(s) \equiv 0$ there is no benefit to be gained from using feedback and the entire discussion makes no sense.

***Only the case $X - BK \equiv 0$ is excluded.

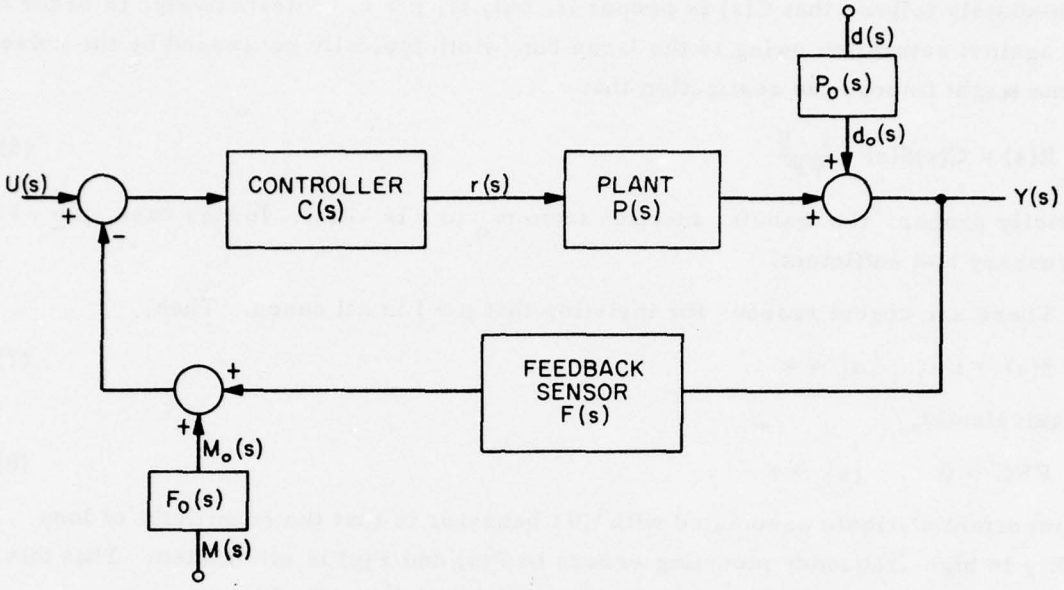


Fig. 1. The system.

The suboptimal design procedure which has been developed produces a controller transfer function which insures that the loop sensitivity function*

$$S(s) = (1 + FPC)^{-1} \tag{2}$$

satisfies**

$$1 - S(s) = O(s^{-\mu}) \quad , \quad |s| \to \infty \tag{3}$$

where μ is a selected positive integer.

The value of μ chosen depends on engineering considerations and design objectives. Suppose, for example, that the restriction C(s) proper is imposed. One easily obtains from Eq. (2) that

$$C(s) = \frac{1 - S}{FPS} \quad . \tag{4}$$

When

---

*Function arguments are omitted wherever convenient.

**$T(s) = O(s^q)$ means that T(s) behaves like s^q for the indicated range of s.

$$FP = O(s^{-\nu}) \quad , \quad |s| \to \infty \quad , \quad \nu > 0 \quad , \tag{5}$$

it immediately follows that C(s) is proper if, only if, $\mu \geq \nu$. Alternatively, in order to guard against saturation owing to the large bandwidth typically possessed by the noise m_o, one might impose the restriction that

$$R(s) = C(s)S(s) = \frac{1 - S}{FP} \tag{6}$$

be strictly proper: the transfer function from m_o to r is -R(s). In this case, $\mu \geq \nu + 1$ is necessary and sufficient.

There are cogent reasons for insisting that $\mu \geq 1$ in all cases. Then,

$$S(s) \to 1 \quad , \quad |s| \to \infty \tag{7}$$

or, equivalently,

$$FPC \to 0 \quad , \quad |s| \to \infty \quad . \tag{8}$$

One important attribute associated with this behavior is that the sensitivity of loop stability to high-frequency modeling errors in P(s) and F(s) is eliminated. That this is the case is readily appreciated with the aid of the Nyquist criterion.

Another factor which enters into the selection of μ is the degree of C(s). It has been established that the larger μ, the higher the degree of C(s). The computations required are least when the smallest μ is selected. In the past, one could also appeal to practical considerations to motivate the choice of C(s) of least possible degree. Today, however, one must[4] recognize "··· the revolutionary impact of the integrated circuit operational amplifier on the practical realizability of compensating transfer functions. Poles and zeros may now be sprinkled like salt and pepper over the s-plane···."

The design procedure described here permits the accommodation of steady-state error specifications. It also allows for including additional design parameters in the numerator polynomial of K(s) which can be selected to improve system performance. For example, design parameters can be introduced to help shape the real-frequency amplitude characteristic of the sensitivity function or to minimize the integral square error. As already pointed out, the characteristic polynomial of the system is selected in advance by the designer and, therefore, the system modes are always under his control. For these reasons, it is believed that the design procedure presented here affords some of the needed flexibility to manage the filter and feedback problems and partially answers the criticisms raised in Ref. 3 concerning the design of single-degree-of-freedom feedback systems.

A. The Design Procedure

The polynomials A(s), B(s), X(s) and Y(s) are obtained from F(s) and P(s). Specifically,

$$F(s)P(s) = \frac{B(s)}{A(s)} \quad , \tag{9}$$

where A(s) and B(s) are relatively prime. It immediately follows that there exist polynomials X(s), Y(s) so that

$$AX + BY \equiv 1 \quad . \tag{10}$$

Let

$$K(s) = \frac{M(s)}{L(s)} \quad . \tag{11}$$

It can be shown that the characteristic polynomial for the closed loop is given by

$$q(s) = h_f h_p h_c \left(\frac{A_f A_p}{A}\right) L \quad , \tag{12}$$

where h_f, h_p, h_c are polynomials which account for the "hidden modes" in the feedback sensor, plant, and controller, respectively. A_f and A_p are the denominator polynomials in the feedback sensor and plant transfer functions, respectively. Clearly, it is necessary that

$$\phi(s) = h_f h_p h_c \left(\frac{A_f A_p}{A}\right) \tag{13}$$

be a strict Hurwitz polynomial* for asymptotic stability and this is assumed to be the case. Also, the controller transfer function can always be obtained with a minimal-order realization so that $h_c(s) \equiv 1$. It is then apparent from Eqs. (12) and (13) that all freedom in the selection of q(s) resides in the choice of L(s). There are no hard and fast rules that can be stated for the selection of L(s). It is at this step in the design that the art of engineering expresses itself. Of course, L(s) strict Hurwitz is necessary and sufficient for asymptotic stability. Also, as established in the sequel, the degree of L(s) is fixed by design objectives.

Since Eq. (3) is to be imposed, it is clear that a useful formula for 1 - S is needed. One easily finds that

$$1 - S = \frac{B(YL + AM)}{L} \quad . \tag{14}$$

---

*One which is free of zeros in the half-plane Re $s \geq 0$.

It is now necessary to distinguish between two cases in order to proceed.

Case 1: FP ≡ constant

In this case, A(s) and B(s) are constants and in Eq. (10) one can choose

$$Y = 0 \quad , \quad X = A^{-1} \quad . \tag{15}$$

Equation (14), then, reduces to

$$1 - S = BA\left(\frac{M}{L}\right) \quad , \tag{16}$$

where the product BA is equal to a nonzero constant. Now Eq. (3) is satisfied provided M and L are chosen so that*

$$\delta(L) = \delta(M) + \mu \quad . \tag{17}$$

As pointed out previously, the design freedom is incorporated into the coefficients of M(s). There are $\delta(M) + 1$ of these and this number is selected by the designer. Then $\delta(L)$ is fixed by Eq. (17) and the zeros of L(s) are picked so that the closed-loop system modes are acceptable. Finally, the coefficients of M(s) are determined so that desired system performance objectives are realized (e.g., a steady-state error specification).

The case described here is not likely in most practical circumstances. In fact, when the plant is nonminimum phase and/or unstable, there is no admissible pair** for which FP is identically constant. Even in the case of minimum-phase asymptotically stable plants, it is not likely that the feedback sensor will cancel all the poles and zeros of the plant. For these reasons, the emphasis in the sequel is on the second case.

Case 2: FP ≢ constant

Either A(s) or B(s) or both are not constant in this case, and it follows that Eq. (7) is satisfied only if

$$\delta(YL) = \delta(AM) \quad . \tag{18}$$

For suppose the contrary is true. Then,

$$\delta(YL + AM) \geq \delta(YL) \geq \delta(L) \quad , \tag{19}$$

and it immediately follows from Eq. (14) that

$$1 - S = O(s^q) \quad , \quad |s| \to \infty \tag{20}$$

---

*For any polynomial $\psi(s)$, $\delta(\psi)$ denotes its degree.

**An admissible pair is one for which $\phi(s)$ in Eq. (13) is strictly Hurwitz.

where

$$q \geq \delta(BY) = \delta(B) + \delta(Y) \quad . \tag{21}$$

Now Eq. (10) requires that

$$\delta(AX) = \delta(BY) \quad . \tag{22}$$

Hence, it is also true that

$$q \geq \delta(A) + \delta(X) \quad . \tag{23}$$

Since $\delta(A) \geq 1$ and/or $\delta(B) \geq 1$, it immediately follows that $q \geq 1$ and Eq. (7) is not satisfied.

Equation (18) is necessary, but not sufficient for the satisfaction of Equation (3). From Eq. (14), it is clear that

$$\delta(L) - \delta(B) - \delta(YL + AM) = \mu \tag{24}$$

is required. In light of Eqs. (18), (24), and $\mu \geq 1$,

$$\delta_r \equiv \delta(YL + AM) < \delta(YL) \equiv n \tag{25}$$

is necessary. Otherwise, the left-hand side of Eq. (24) is

$$\delta(L) - \delta(B) - \delta(YL) = -\delta(B) - \delta(Y) \leq 0 \tag{26}$$

and $\mu \geq 1$ is impossible. Substituting Eq. (25) into Eq. (24) yields

$$\delta_r = \delta(L) - \delta(B) - \mu \tag{27}$$

as the required value for the reduced degree of YL + AM. It is later verified that the designer can always guarantee $\delta(L) \geq \delta(B) + \mu$ so that $\delta_r \geq 0$ as it must be. Attention is now turned to the selection of M(s) so that the polynomial YL + AM has degree $\delta_r < n$.

Let

$$YL = \sum_{k=0}^{n} \gamma_k s^k \quad , \tag{28}$$

$$A = \sum_{i=0}^{n_a} a_i s^i \quad , \tag{29}$$

and

$$M = \sum_{j=0}^{n_m} m_j s^j \quad . \tag{30}$$

Then, in view of Eq. (18),

$$AM = \sum_{i=0}^{n_a} \sum_{j=0}^{n_m} a_i m_j s^{i+j} = \sum_{k=0}^{n} c_k s^k \quad , \tag{31}$$

where

$$c_k = \sum_{j=0}^{k} a_{k-j} m_j U(n_m - j)\, U(n_a - k + j) \tag{32}$$

and

$$U(q) \equiv \begin{cases} 1, & q \geq 0 \\ 0, & q < 0 \end{cases} \quad . \tag{33}$$

The conditions, hence, which assure that the polynomial YL + AM has degree δ_r are

$$c_k = -\gamma_k \quad , \quad k = (\delta_r + 1) \rightarrow n \quad . \tag{34}$$

Since,

$$n_a - k + j \geq 0 \tag{35}$$

only if

$$j \geq k - n_a \quad , \tag{36}$$

it readily follows from Eqs. (32) and (34) that

$$\sum_{j=k-n_a}^{k} a_{k-j} m_j U(n_m - j) = -\gamma_k \quad , \quad k = (\delta_r + 1) \rightarrow n \quad . \tag{37}$$

Equation (37) constitutes a system of $n - \delta_r$ linear inhomogeneous equations in the

$$n_m - (\delta_r + 1 - n_a) + 1 = n_m + n_a - \delta_r = n - \delta_r \tag{38}$$

unknowns

$$m_{n_m}, m_{n_m - 1}, \cdots, m_{\delta_r + 1 - n_a} \quad . \tag{39}$$

It is easily shown that this system of equations always has a unique solution for the unknowns provided

$$\delta_f \equiv \delta_r + 1 - n_a = \delta_r + 1 - \delta(A) \geq 0 \quad . \tag{40}$$

Note that δ_f is the number of remaining coefficients in the polynomial M(s) which

can be used to satisfy design objectives other than Equation (3). Once δ_f is chosen, Eq. (40) fixes the value of δ_r and it follows from Eqs. (27) and (40) that

$$\delta(L) = \delta(B) + \mu + \delta_f + \delta(A) - 1 \quad . \tag{41}$$

Equation (41) gives the degree of L(s) in terms of known data once μ and δ_f have been selected. Clearly, the condition $\delta(L) \geq \delta(B) + \mu$ imposed by Eq. (27) is satisfied whenever

$$\delta_f + \delta(A) \geq 1 \quad . \tag{42}$$

Since almost all problems of interest require some design freedom, $\delta_f \geq 1$, and Eq. (42) is satisfied. It is also rare to find any practical cases in which $\delta(A) = 0$. Hence, there is little, if any, real loss in imposing the restriction, Eq. (42), on the choice of δ_f, and this is assumed in the sequel.

The degree of the polynomial M(s) is fixed by Eq. (18). Specifically,

$$\delta(M) = \delta(Y) + \delta(L) - \delta(A) \tag{43}$$

or, with the aid of Eq. (41),

$$\delta(M) = \delta(Y) + \delta(B) + \mu + \delta_f - 1 \quad . \tag{44}$$

Equations (41) and (44) are key relationships. The first step in the design procedure is to select μ and δ_f. Then Eq. (41) fixes the degree of L(s). The zeros of L(s) are chosen next keeping in mind that L(s) is a factor in the characteristic polynomial for the closed loop. Once L(s) has been picked and Y(s) has been selected so that Eq. (10) is satisfied,* the degree of M(s) is set by Eq. (43). The first $n - \delta_r$ coefficients of M(s) are then found by solving Equations (37). The remaining δ_f coefficients of M(s) are set by design specifications. Once L and M are known, a straightforward computation gives C(s) from

$$C(s) = \frac{YL + AM}{XL - BM} = \frac{B_c}{A_c} \quad . \tag{45}$$

When μ is selected so that the controller transfer function is proper or strictly proper, then the minimal order of the dynamical system which realizes this transfer function is the degree of $A_c(s)$, the denominator of C(s). It is not difficult to show that

$$S = 1 - (1 - S) = \frac{(XL - BM)A}{L} = \frac{A_c A}{L} \quad . \tag{46}$$

It immediately follows that

---

*There exist unique polynomials X(s), Y(s) satisfying Eq. (10) and $\delta(X) \leq \delta(B) - 1$, $\delta(Y) \leq \delta(A) - 1$. See Ref. 5, Theorem 4.

$$\delta(A_c) = \delta(L) - \delta(A) = \delta(B) + \mu + \delta_f - 1 \tag{47}$$

since $S \to 1$, $|s| \to \infty$. Equation (46) is of interest in those cases where the degree of C(s) is an important design factor.

When C(s) is improper, not only is the degree of the denominator important, but the degree of the numerator as well. It follows from Eqs. (45), (25) and (40) that

$$\delta(B_c) = \delta_r = \delta_f + \delta(A) - 1 \quad . \tag{48}$$

It is interesting (and comforting) to note that both $\delta(A_c)$ and $\delta(B_c)$ are independent of the choice of X(s), Y(s) satisfying Equation (10). That this is the case is obvious upon inspection of Eqs. (47) and (48).

It is possible for the computed M(s) not to be relatively prime to the selected L(s). A factor common to both YL + AM and XL - BM then exists. When the degree of this factor is denoted by δ_c, one obtains in place of Eqs. (47) and (48)

$$\delta(A_c) = \delta(B) + \mu + \delta_f - 1 - \delta_c \tag{49}$$

and

$$\delta(B_c) = \delta_f + \delta(A) - 1 - \delta_c \quad , \tag{50}$$

respectively. It is also true in this case that Eq. (12) is no longer the characteristic polynomial for the closed loop. The correct expression is obtained from Eq. (12) by replacing L(s) with the polynomial which results when L(s) is divided by the factor common to M and L. It is recommended that all designs be initiated under the assumption that L and M turn out to be relatively prime: this is the most likely case.

B. Discussion

A straightforward design procedure for single-input-output systems described by rational transfer functions has been developed. The method has been applied to several examples and these details are contained in a paper to be submitted for publication. The design method is easily programmed on a digital computer and should replace all classical design techniques based on the root locus method. The fact that the closed-loop system modes are selected by the designer at the outset is a significant advantage of the approach described here. The ability to introduce the constraints Eqs. (3) and (7) is important practically and this cannot be overemphasized.

National Science Foundation
ENG74-13054

J.J. Bongiorno, Jr.

AD-A034 764 POLYTECHNIC INST OF NEW YORK BROOKLYN MICROWAVE RESE--ETC F/G 20/3
PROGRESS REPORT NUMBER 41, 15 SEPTEMBER 1975 THROUGH 14 SEPTEMB--ETC(U)
NOV 76 A A OLINER F44620-74-C-0056

UNCLASSIFIED POLY-MRI-452.41-76 NL

7 OF 7
AD
A034764



END

DATE
FILMED
3-77

REFERENCES

1. D.C. Youla, J.J. Bongiorno, Jr. and H. Jabr, "Modern Wiener-Hopf Design of Optimal Controllers - Part I: The Single-Input-Output Case," IEEE Trans. on Automatic Control, AC-21, No. 1, 3-13 (February 1976).

2. D.C. Youla, H. Jabr and J.J. Bongiorno, Jr., "Modern Wiener-Hopf Design of Optimal Controllers - Part II: The Multivariable Case," IEEE Trans. on Automatic Control, AC-21, No. 3, 319-338 (June 1976).

3. I.M. Horowitz, "Synthesis of Feedback Systems," Sec. 5.17 (New York: Academic Press, 1963).

4. D.P. Petersen, Book Reviews Section of "Circuits and Systems," 5-6 (June 1976).

5. M. Bôcher, "Introduction of Higher Algebra," Sec. 67 (New York: Macmillan, 1907).

GENERALIZED IMAGE RESTORATION BY THE METHOD OF ALTERNATING ORTHOGONAL PROJECTIONS

D.C. Youla

The view adopted in this report is that the problem of image restoration is essentially geometric in character and can be formulated as follows: The complete image f is a vector known a priori to belong to a linear subspace $\mathcal{Q}_b$ but all that is available is its projection $P_a f$ onto a known linear subspace $\mathcal{Q}_a$. 1) Find the necessary and sufficient conditions under which f is uniquely determined by $P_a f$ and 2) find the necessary and sufficient conditions guaranteeing the stable reconstruction of f from $P_a f$ in the face of noise. (In the latter case the reconstruction problem is said to be well-posed.) The answers turn out to be remarkably simple.

(a_1) f is uniquely determined by $P_a f$ iff $\mathcal{Q}_b$ and the orthogonal complement of $\mathcal{Q}_a$ have only the zero vector in common.

(a_2) The reconstruction problem is well-posed iff the angle between $\mathcal{Q}_b$ and the orthogonal complement of $\mathcal{Q}_a$ is greater than zero. (All angles lie in the first quadrant.)

(a_3) In both cases (a_1) and (a_2) there exists an effective recursive algorithm for the recovery of f employing only the operations of projection onto $\mathcal{Q}_b$ and projection onto the orthogonal complement of $\mathcal{Q}_a$.

It appears natural to carry out the proofs of (a_1) - (a_3) in a Hilbert space setting.

A. Preliminaries*

Consider a Hilbert space $\mathcal{H}$ with elements f, g, h, x, z, etc., a zero vector ϕ and an inner product (x, y). By definition,

$$\|f\| = \sqrt{(f,f)} \geq 0 \tag{1}$$

is the "length" of f and the sequence $\{f_k\}$ is said to converge to $f(f_k \to f)$ if

$$\lim_{k\to\infty} \|f_k - f\| = 0 \quad . \tag{2}$$

Let $\mathcal{Q}$ be any closed linear manifold (CLM) in $\mathcal{H}$ and $\perp\mathcal{Q}$ its orthogonal complement. According to the projection theorem[1] every $f \in \mathcal{H}$ possesses a unique decomposition

$$f = g + h \tag{3}$$

in which $g \in \mathcal{Q}$ and $h \in \perp\mathcal{Q}$. Symbolically,

---

*All the general functional analysis needed to read this section can be found in Reference 1. Regarding notation, the symbols $\in$, $\cap$, $\cup$ and 1 stand for set membership, intersection, union and the identity operator. A closed linear manifold is a linear manifold which contains all its limit points (and therefore equals its closure). Our Hilbert space $\mathcal{H}$ is assumed to be complete.

$$\mathcal{H} = \mathcal{\theta} \dot{+} (\perp \mathcal{\theta}) \quad . \tag{4}$$

Since $(g,h) = 0$, the vectors g and h are mutually orthogonal and the two linear operators P and Q defined by the rules $g = Pf$, $h = Qf$ are the associated orthogonal projection operators projecting onto $\mathcal{\theta}$ and $\perp\mathcal{\theta}$, respectively. As is well known,[1] $P^2 = P$ and $Q^2 = Q = 1 - P$. Moreover, P is self-adjoint ($P = P^*$) whence $(Px, y) = (x, Py)$ for all x and y in $\mathcal{H}$.

Now suppose an element $f \varepsilon \mathcal{H}$ belongs to a known CMP $\mathcal{\theta}_b$ but we are given only its projection* $g = P_a f$ onto the known CLM $\mathcal{\theta}_a$. How do we go about reconstructing f from g? Let P_a, Q_a, P_b, Q_b denote the projection operators projecting onto $\mathcal{\theta}_a$, $\perp\mathcal{\theta}_a$, $\mathcal{\theta}_b$ and $\perp\mathcal{\theta}_b$, respectively. Then, $f \varepsilon \mathcal{\theta}_b$ implies $f = P_b f$ and

$$g = P_a f = P_a P_b f = (1 - Q_a) P_b f = P_b f - Q_a P_b f = f - Q_a P_b f \quad . \tag{5}$$

Thus f satisfies the operator equation

$$Af = g \tag{6}$$

where

$$A = 1 - Q_a P_b \quad . \tag{7}$$

Clearly, A is a bounded** operator whose domain D(A) is all of $\mathcal{H}$ and whose range R(A) is the set of all vectors $g \varepsilon \mathcal{H}$ admitting the representation Eq. (6) for some choice of $f \varepsilon \mathcal{H}$.

The vector f is uniquely determined by g iff the inverse operator

$$T = A^{-1} \tag{8}$$

exists for then

$$f = Tg \quad . \tag{9}$$

However, to examine the matter of stable reconstruction it is necessary to evaluate the performance of Eq. (9) in the presence of noise. The effect of noise is to corrupt g to

---

*Unless stated explicitly otherwise, all projections are orthogonal.

**A linear operator T is bounded if

$$\|T\| \equiv \sup_{x \varepsilon D(T)} \frac{\|Tx\|}{\|x\|} < \infty \quad . \tag{9a}$$

Any orthogonal projection operator has norm equal to one.

$g + \Delta g$ and

$$\Delta f = T(\Delta g) \tag{10}$$

is the error induced in f. In general there is no restriction on the orientation of Δg in $\mathcal{H}$ and the reconstruction process Eq. (9) is executable without exception iff the domain of T is all of $\mathcal{H}$. Moreover, to preclude the possibility of arbitrarily large percentage errors one must also impose the condition

$$\sup_{\Delta g \in \mathcal{H}} \frac{\|T(\Delta g)\|}{\|\Delta g\|} < \infty \quad . \tag{11}$$

In other words,

$$\|T\| = \|A^{-1}\| < \infty \tag{12}$$

and the inverse operator is forced to be bounded. The task is to translate the above requirements onto θ_a and θ_b.

The algorithm for the recovery of f is inspired by the equation

$$f = g + Q_a P_b f \tag{13}$$

which immediately suggests the fundamental recursion

$$f_{k+1} = g + Q_a P_b f_k \quad , \quad k = 1 \to \infty \; ; \; f_1 = g \quad . \tag{14}$$

Read from right to left Eq. (14) states that f_{k+1} is created by projecting f_k onto θ_b, then projecting the result onto $\perp \theta_a$ and finally adding g to restore the correct projection onto θ_a. The geometric significance of this cycle of operations can be grasped very easily with the help of the two diagrams in Figures 1(a) and 1(b).

The three CLMS θ_a, $\perp \theta_a$ and θ_b are indicated as three straight lines of infinite extent passing through the origin. Although the instrumentality is incapable of synthesizing f from $P_a f = OB = g$ by direct movement back along the perpendicular BA to θ_a, it is nevertheless possible to reach intermediate points such as D, F, H, etc., which tend to the limit A. For starting with OB and projecting it onto θ_b we obtain OC. Since the projection of OC onto θ_a is unequal to OB and therefore incorrect, the next step is one of restoration. This is accomplished by projecting OC onto $\perp \theta_a$ and then adding the result OC″(=BD) to OB. Thus OD is the second approximation to f. Further repetitions of this method of alternating projections yield the length-increasing vector approximations OF, OH, etc., which obviously converge to f = OA provided the lines $\perp \theta_a$ and θ_b do not coincide. We now proceed to show that this heuristic reasoning can be made rigorous.

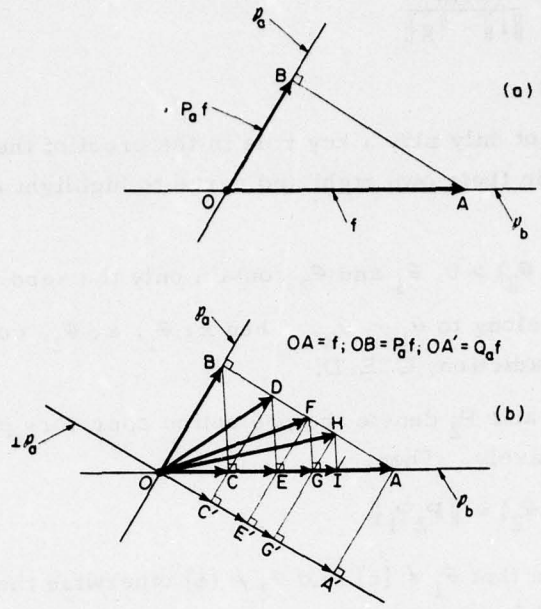


Fig. 1. The geometry of reconstruction in Hilbert space.

B. The Main Results

The inner product permits us to define the angle between two elements f and g in $\mathcal{H}$ as follows. By the Schwartz inequality we know that

$$0 \leq \frac{|(f,g)|}{\|f\| \cdot \|g\|} \leq 1 \quad . \tag{15}$$

We may therefore define the angle $\psi(f,g)$ between f and g by

$$\cos \psi(f,g) = \frac{|(f,g)|}{\|f\| \cdot \|g\|} \quad , \quad 0 \leq \psi(f,g) \leq \frac{\pi}{2} \quad . \tag{16}$$

The angle $\psi(\theta_1, \theta_2)$ between two linear manifolds θ_1 and θ_2 is defined by the expression*

$$\psi(\theta_1, \theta_2) = \inf_{\substack{f \in \theta_1 \\ g \in \theta_2}} \psi(f,g) \quad . \tag{17}$$

Equivalently, since cos ψ is a monotone decreasing function of ψ in the interval

---

*If either θ_1 or $\theta_2 = \{\phi\}$, $\cos \psi(\theta_1, \theta_2) \equiv 0$ and $\psi(\theta_1, \theta_2) = 90^\circ$.

$0 \leq \psi \leq \pi/2$,

$$\cos \psi(\mathcal{G}_1, \mathcal{G}_2) = \sup_{\substack{f \in \mathcal{G}_1 \\ g \in \mathcal{G}_2}} \frac{|(f,g)|}{\|f\| \cdot \|g\|} \quad . \tag{18}$$

The next three lemmas not only play a key role in the proof of theorem 1 but are also of considerable interest in their own right and serve to highlight the importance of the angle concept.

Lemma 1. If $\psi(\mathcal{G}_1, \mathcal{G}_2) > 0$, $\mathcal{G}_1$ and $\mathcal{G}_2$ contain only the zero vector ϕ in common.

Proof. Let $x \neq \phi$ belong to $\mathcal{G}_1 \cap \mathcal{G}_2$. Then $x \in \mathcal{G}_1$, $x \in \mathcal{G}_2$, $\cos \psi(x,x) = 1$, $\psi(x,x) = 0$ and $\psi(\mathcal{G}_1, \mathcal{G}_2) = 0$, a contradiction, Q.E.D.

Lemma 2. Let P_1 and P_2 denote the projection operators projecting onto the CLMS $\mathcal{G}_1$ and $\mathcal{G}_2$, respectively. Then,

$$\|P_1 P_2\| = \cos \psi(\mathcal{G}_1, \mathcal{G}_2) = \|P_2 P_1\| \quad . \tag{19}$$

Proof. (We assume that $\mathcal{G}_1 \neq \{\phi\}$ and $\mathcal{G}_2 \neq \{\phi\}$ otherwise the lemma is trivial.) For any $f \in \mathcal{G}_1$, $f = P_1 f$ and for fixed $g \in \mathcal{G}_2$,

$$\frac{|(f,g)|}{\|f\| \cdot \|g\|} = \frac{|(P_1 f, g)|}{\|f\| \cdot \|g\|} = \frac{|(f, P_1 g)|}{\|f\| \cdot \|g\|} \leq \frac{\|P_1 g\|}{\|g\|} \tag{20}$$

with equality iff $f = \lambda P_1 g \in \mathcal{G}_1$, λ an arbitrary scalar. Thus, since $g = P_2 g$,

$$\cos \psi(\mathcal{G}_1, \mathcal{G}_2) = \sup_{g \in \mathcal{G}_2} \frac{\|P_1 P_2 g\|}{\|g\|} \quad . \tag{21}$$

From the definition Eq. (9a) of operator norm it is clear that the sup on the right-hand side of Eq. (21) cannot exceed $\|P_1 P_2\|$. Conversely, if h is an arbitrary member of $\mathcal{H}$, $h = g + x$ where $g \in \mathcal{G}_2$ and $x \in \perp \mathcal{G}_2$ and

$$\frac{\|P_1 P_2 h\|}{\|h\|} = \frac{\|P_1 P_2 g\|}{\|h\|} = \frac{\|P_1 P_2 g\|}{(\|g\|^2 + \|x\|^2)^{1/2}} \leq \frac{\|P_1 P_2 g\|}{\|g\|} \tag{22}$$

$$\therefore \sup_{g \in \mathcal{G}_2} \frac{\|P_1 P_2 g\|}{\|g\|} \leq \|P_1 P_2\| \leq \sup_{g \in \mathcal{G}_2} \frac{\|P_1 P_2 g\|}{\|g\|} \quad . \tag{23}$$

implying

$$\|P_1P_2\| = \sup_{g\varepsilon\mathcal{G}_2} \frac{\|P_1P_2g\|}{\|g\|} = \sup_{g\varepsilon\mathcal{G}_2} \frac{\|P_1g\|}{\|g\|} \quad . \tag{24}$$

Since $\|P_1P_2\| = \|(P_1P_2)^*\| = \|P_2P_1\|$, Eq. (19) is established.

Lemma 3. Let $\mathcal{G}_1$ and $\mathcal{G}_2$ be two closed linear manifolds such that $\mathcal{G}_1 \cap \mathcal{G}_2 = \phi$. Then, $\mathcal{G} = \mathcal{G}_1 + \mathcal{G}_2$ is closed if and only if $\psi(\mathcal{G}_1, \mathcal{G}_2) > 0$.

Proof.* Sufficiency. According to lemma 2, $\psi(\mathcal{G}_1, \mathcal{G}_2) > 0$ implies $\|P_2P_1\| < 1$, P_1 and P_2 the projection operators projecting onto $\mathcal{G}_1$ and $\mathcal{G}_2$, respectively. Recall that $\mathcal{G}$ is the set of all vectors of the form $f = x+y$, $x\varepsilon\mathcal{G}_1$, $y\varepsilon\mathcal{G}_2$. Let $\{f_k = x_k + y_k\}$ be any sequence in $\mathcal{G}$ converging to a limit f. What must be shown is that $f\varepsilon\mathcal{G}$. Write

$$f_k = (x_k - P_2x_k) + (y_k + P_2x_k) \quad , \tag{25}$$

$x_k\varepsilon\mathcal{G}_1$, $y_k\varepsilon\mathcal{G}_2$, $k = 1 \to \infty$. Since $x_k - P_2x_k\varepsilon\perp\mathcal{G}_2$ and $y_k + P_2x_k\varepsilon\mathcal{G}_2$,

$$\|f_k - f_\ell\|^2 = \|(1 - P_2P_1)(x_k - x_\ell)\|^2 + \|(y_k - y_\ell) + P_2P_1(x_k - x_\ell)\|^2 \quad , \tag{26}$$

$k, \ell = 1 \to \infty$. Hence $\|f_k - f_\ell\| \to 0$ implies

$$\|(1 - P_2P_1)(x_k - x_\ell)\| \to 0 \tag{27}$$

and

$$\|(y_k - y_\ell) + P_2P_1(x_k - x_\ell)\| \to 0 \quad . \tag{28}$$

But

$$\|(1 - P_2P_1)(x_k - x_\ell)\| \geq (1 - \|P_2P_1\|) \cdot \|x_k - x_\ell\| \tag{29}$$

which together with $\|P_2P_1\| < 1$ and Eq. (27) yields $\|x_k - x_\ell\| \to 0$. Thus $x_k \to x\varepsilon\mathcal{G}_1$ because $\mathcal{G}_1$ is closed. Invoking Eq. (28), $\|y_k - y_\ell\| \to 0$ whence $y_k \to y\varepsilon\mathcal{G}_2$ because $\mathcal{G}_2$ is closed. It now follows from $f_k = x_k + y_k$ that $f = x+y\varepsilon\mathcal{G}$, Q.E.D.

Necessity (the difficult part). Suppose $\mathcal{G} = \mathcal{G}_1 + \mathcal{G}_2$ is closed. Then $\mathcal{G}$, considered by itself is a complete Hilbert space. In view of the assumption $\mathcal{G}_1 \cap \mathcal{G}_2 = \{\phi\}$, $\mathcal{G}$ actually equals $\mathcal{G}_1 \dot{+} \mathcal{G}_2$, the "direct" sum of $\mathcal{G}_1$ and $\mathcal{G}_2$. Every vector f in $\mathcal{G}$ therefore possesses a unique decomposition $f = x+y$, $x\varepsilon\mathcal{G}_1$, $y\varepsilon\mathcal{G}_2$ and the rule $x = Lf$ defines a linear (oblique) projection operator L projecting $\mathcal{G}$ onto $\mathcal{G}_1$ parallel to $\mathcal{G}_2$. The domain of L is all of $\mathcal{G}$ and to prove that it is bounded in $\mathcal{G}$, i.e., that

---

*We assume that $\mathcal{G}_1 \neq \{\phi\}$ and $\mathcal{G}_2 \neq \{\phi\}$ otherwise the lemma is trivial.

$$\|L\|_{\mathcal{\theta}} \equiv \sup_{f\varepsilon\mathcal{\theta}} \frac{\|Lf\|}{\|f\|} < \infty \quad , \tag{30}$$

it suffices to establish that it is closed in $\mathcal{\theta}$ (Ref. 1, p. 165, Thm. 8). Now for every $f\varepsilon\mathcal{\theta}$ we have $f = Lf + (f - Lf)$. Let $\lim f_k = f_o$, $\lim Lf_k = g_o$, $f_k \varepsilon \mathcal{\theta}$, $k = 1 \to \infty$. To verify that L is closed in $\mathcal{\theta}$ we must show that $g_o = Lf_o$. Since all f_k are in $\mathcal{\theta}$ and all Lf_k are in $\mathcal{\theta}_1$ and all $f_k - Lf_k$ are in $\mathcal{\theta}_2$, we have $f_o \varepsilon \mathcal{\theta}$, $g_o \varepsilon \mathcal{\theta}_1$ and $\lim(f_k - Lf_k) = f_o - g_o = h_o \varepsilon \mathcal{\theta}_2$ because $\mathcal{\theta}$, $\mathcal{\theta}_1$ and $\mathcal{\theta}_2$ are closed. Consequently, $f_o = g_o + h_o$, $g_o \varepsilon \mathcal{\theta}_1$, $h_o \varepsilon \mathcal{\theta}_2$ and it follows that $g_o = Lf_o$; i.e., L is closed in $\mathcal{\theta}$, Q.E.D.

In this final step of the proof we relate $\|L\|_{\mathcal{\theta}}$ to $\sin \psi(\mathcal{\theta}_1, \mathcal{\theta}_2)$. Clearly, since any f in $\mathcal{\theta}$ is of the form $f = x+y$, $x \varepsilon \mathcal{\theta}_1$, $y \varepsilon \mathcal{\theta}_2$,

$$\|L\|_{\mathcal{\theta}} = \sup_{f\varepsilon\mathcal{\theta}} \frac{\|Lf\|}{\|f\|} = \sup_{\substack{x\varepsilon\mathcal{\theta}_1 \\ y\varepsilon\mathcal{\theta}_2}} \frac{\|x\|}{\|x+y\|} \quad . \tag{31}$$

With $x \varepsilon \mathcal{\theta}_1$ held fixed,

$$\inf_{y\varepsilon\mathcal{\theta}_2} \|x+y\| = \text{distance from x to } \mathcal{\theta}_2 = \|Q_2 x\| \tag{32}$$

and using $x = P_1 x$ Eq. (31) goes into

$$\|L\|_{\mathcal{\theta}} = \sup_{x\varepsilon\mathcal{\theta}_1} \left(\frac{\|Q_2 P_1 x\|}{\|x\|}\right)^{-1} = \left(\inf_{x\varepsilon\mathcal{\theta}_1} \frac{\|Q_2 P_1 x\|}{\|x\|}\right)^{-1} \quad .$$

Writing $P_1 x = P_2 P_1 + Q_2 P_1 x$ we have $\|P_1 x\|^2 = \|P_2 P_1 x\|^2 + \|Q_2 P_1 x\|^2$ from which

$$\sup_{x\varepsilon\mathcal{\theta}_1} \left(1 - \frac{\|Q_2 P_1 x\|^2}{\|P_1 x\|^2}\right) = \left(1 - \inf_{x\varepsilon\mathcal{\theta}_1} \frac{\|Q_2 P_1 x\|^2}{\|x\|^2}\right) = \sup_{x\varepsilon\mathcal{\theta}_1} \frac{\|P_2 P_1 x\|^2}{\|x\|^2} = \cos^2 \psi(\mathcal{\theta}_1, \mathcal{\theta}_2) \ ,$$

by lemma 2.

$$\therefore \quad \|L\|_{\mathcal{\theta}} = \frac{1}{\sin \psi(\mathcal{\theta}_1, \mathcal{\theta}_2)} \tag{33}$$

and $\|L\|_{\mathcal{\theta}} < \infty$ implies $\psi(\mathcal{\theta}_1, \mathcal{\theta}_2) > 0$. This completes the proof of lemma 3, Q.E.D.

Theorem 1. Let $\mathcal{\theta}_a$ and $\mathcal{\theta}_b$ be any two closed linear manifolds in $\mathcal{H}$ and P_a, Q_a, P_b, Q_b (orthogonal) projection operators projecting onto $\mathcal{\theta}_a$, $\perp\mathcal{\theta}_a$, $\mathcal{\theta}_b$ and $\perp\mathcal{\theta}_b$, respectively. Suppose $f \varepsilon \mathcal{\theta}_b$. Then,

(a_1) f is uniquely determined by its projection $P_a f$ onto $\mathcal{\theta}_a$ if and only if

$$\mathcal{\theta}_b \cap (\perp \mathcal{\theta}_a) = \{\phi\} \quad . \tag{34}$$

(a_2) The problem of reconstructing f from P_a is well-posed if and only if

$$\psi(\mathcal{\theta}_b, \perp \mathcal{\theta}_a) > 0 \quad . \tag{35}$$

This angle constraint is satisfied iff

$$\| Q_a P_b \| \equiv \rho < 1 \quad . \tag{35a}$$

(a_3) In both cases (a_1) and (a_2), the sequence $\{f_k\}$ generated by the recursion

$$f_{k+1} = g + Q_a P_b f_k \quad , \quad k = 1 \to \infty \; ; \; f_1 = g = P_a f \tag{36}$$

converges to f in norm, i.e.,

$$\lim_{k \to \infty} f_k = f \quad . \tag{37}$$

The convergence is strict monotone increasing; i.e., $f_k \uparrow f$.

Proof. (a_1) According to Eqs. (6) and (7), $Af = g$ where $A = 1 - Q_a P_b$. The vector f is uniquely determined by g iff A^{-1} exists; i.e., iff the equation

$$Af_o = (1 - Q_a P_b) f_o = \phi$$

possesses the unique solution $f_o = \phi$. Clearly, $f_o = Q_a P_b f_o$ implies $\| f_o \| = \| Q_a P_b f_o \|$ which is possible iff $f_o \in \mathcal{\theta}_b \cap (\perp \mathcal{\theta}_a)$. Conversely, any $f_o \in \mathcal{\theta}_b \cap (\perp \mathcal{\theta}_a)$ satisfies $Af_o = \phi$ whence, A^{-1} exists iff $\mathcal{\theta}_b \cap (\perp \mathcal{\theta}_a) = \{\phi\}$, Q.E.D.

(a_2) If the reconstruction problem is well-posed $T = A^{-1}$ must exist and be bounded. The domain D(T) of T is the range of A which is the collection of all vectors g of the form

$$g = f - Q_a P_b f \quad , \quad f \in \mathcal{H} \quad . \tag{38}$$

The existence of T implies that D(T) is dense in $\mathcal{H}$. Suppose to the contrary that some $h \in \mathcal{H}$ is orthogonal to D(T). Then,

$$0 = (h, Af) = (A^* h, f) = (1 - P_b Q_a) h, f)$$

for all $f \in \mathcal{H}$. Hence $(1 - P_b Q_a)h = \phi$ or $h = P_b Q_a h$ and reasoning as before $h \in \mathcal{\theta}_b \cap (\perp \mathcal{\theta}_a)$. From ($a_1$), $h = \phi$ proving that D(T) is dense in $\mathcal{H}$.

Let us now impose the additional condition that T be bounded and let g be an arbitrary element in $\mathcal{H}$. Since D(T) is dense in $\mathcal{H}$ there exists a sequence

$$g_k = f_k - Q_a P_b f_k \quad , \quad k = 1 \to \infty \quad , \tag{39}$$

which is contained in D(T) and converges to g. From $f_k = Tg_k$ and the boundedness of T we conclude that the sequence $\{f_k\}$ also converges to some $f \varepsilon \mathcal{H}$. Going to the limit in Eq. (39),

$$g = f - Q_a P_b f = (1 - Q_a P_b) f \quad . \tag{40}$$

Thus every $g \varepsilon R(A)$ and the domain of T is the entire space $\mathcal{H}$. Writing $f = P_b f + Q_b f$, it follows from Eq. (40) that every $g \varepsilon \mathcal{H}$ admits the representation

$$g = Q_b f + P_a P_b f = x + y \tag{41}$$

where $x = Q_b f \varepsilon \perp \mathcal{G}_b$ and $y = P_a P_b f \varepsilon \mathcal{G}_a$. In other words, $(\perp \mathcal{G}_a) + \mathcal{G}_a = \mathcal{H}$, a CLM. But then (Ref. 2, Thm. 4.8), $\perp(\perp \mathcal{G}_b) + \perp \mathcal{G}_a = \mathcal{G}_b \dot{+} \perp \mathcal{G}_a$ is also closed and invoking lemmas 3 and 2, $\psi(\mathcal{G}_b, \perp \mathcal{G}_a) > 0$ or,

$$\rho \equiv \|Q_a P_b\| = \cos \psi(\mathcal{G}_b, \perp \mathcal{G}_a) < 1 \quad . \tag{42}$$

Inversely, if $\psi(\mathcal{G}_b, \perp \mathcal{G}_a) > 0$, $\mathcal{G}_b \cap (\perp \mathcal{G}_a) = \{\phi\}$ by lemma 1 and the explicit solution of Eq. (40) is given by

$$f = \sum_{r=o}^{\infty} (Q_a P_b)^r g \tag{43}$$

which converges like a geometric series since $\|Q_a P_b\| = \rho < 1$. From Eq. (41), if $f \varepsilon \mathcal{G}_b$ then $g = P_a P_b f \varepsilon \mathcal{G}_a$. Again, if $g \varepsilon \mathcal{G}_a$, $Q_a Q_b f = 0$ implying that $Q_b f$ belongs to both $\perp \mathcal{G}_b$ and $\mathcal{G}_a$. Thus, if $(\perp \mathcal{G}_b) \cap \mathcal{G}_a = \{\phi\}$, $Q_b f = \{\phi\}$ and $f \varepsilon \mathcal{G}_b$. Consequently, under the additional (but unnecessary) assumption $(\perp \mathcal{G}_b) \cap \mathcal{G}_a = \{\phi\}$, $f \varepsilon \mathcal{G}_b$ iff $g \varepsilon \mathcal{G}_a$. Finally, from Eq. (40)

$$\|f\| \leq \frac{1}{1 - \|Q_a P_b\|} \cdot \|g\|$$

which coupled with Eq. (42) yields

$$\|A^{-1}\| \leq \frac{1}{1 - \cos \psi(\mathcal{G}_b, \perp \mathcal{G}_a)} \quad . \tag{44}$$

(a_3) If g is in the range of A it has the form $g = f - Q_a P_b f$. Iterating Eq. (36) we find that

$$f_k = \sum_{r=o}^{k-1} (Q_a P_b)^r g = \sum_{r=o}^{k-1} (Q_a P_b)^r (f - Q_a P_b f) = f - (Q_a P_b)^k f \quad . \tag{45}$$

According to von Neumann's theorem on alternating projections (Ref. 3, Thm. 13.7), for every $f \psi \mathcal{H}$

$$\lim_{k\to\infty} (Q_a P_b)^k f = f_c \tag{46}$$

where f_c is the projection of f onto the CLM $\mathcal{Q}_c = \mathcal{Q}_b \cap (\perp \mathcal{Q}_a)$. Since $\mathcal{Q}_c = \{\phi\}$, $f_c = \phi$ and $\lim f_k = f$ in both cases (a_1) and (a_2). That $f_k \uparrow f$ is easily seen from Eq. (45). The error

$$f - f_k = \varepsilon_k = (Q_a P_b)^k f \quad , \; k = 1 \to \infty \quad . \tag{47}$$

Clearly, $\varepsilon_{k+1} = (Q_a P_b)\varepsilon_k$ and $\|\varepsilon_{k+1}\| \leq \|\varepsilon_k\|$ with equality iff $\varepsilon_k \varepsilon \mathcal{Q}_b \cap (\perp \mathcal{Q}_a)$, i.e., iff $\varepsilon_k = \phi$. It follows that if k_o is the smallest integer for which $\|\varepsilon_{k+1}\| = \|\varepsilon_k\|$, then $\varepsilon_k = \phi$, $k \geq k_o$ and $\|\varepsilon_{k+1}\| < \|\varepsilon_k\|$, $k < k_o$; i.e., $f_k \uparrow f$ and theorem 1 is proved, Q.E.D.

As we have seen, if $\psi(\mathcal{Q}_b, \perp \mathcal{Q}_a) = 0$ the reconstruction problem is unstable ($\|A^{-1}\| = \infty$) and the numerical performance of Eq. (36) depends on the fine details of the noise and its specific "orientation" with respect to the CLMS $\mathcal{Q}_b$ and $\perp \mathcal{Q}_a$. However, if $\psi(\mathcal{Q}_b, \perp \mathcal{Q}_a) > 0$ a satisfactory error analysis is possible.

Corollary 1. Let noise corrupt $g = P_a f$ to $g + \Delta g$ and let $\{f_k\}$ denote the sequence of approximations generated by the recursion Eq. (36) under the initialization $f_1 = g + \Delta g$. Let $f_o = \lim f_k$ and $f_o - f = \Delta f$. Suppose $\psi(\mathcal{Q}_b, \perp \mathcal{Q}_a) > 0$. Then,

(1) if $f \varepsilon \mathcal{Q}_b$, $\Delta g \varepsilon \mathcal{Q}_a$ and $(\perp \mathcal{Q}_b) \cap \mathcal{Q}_a = \{\phi\}$,

$$\|\Delta f\| \leq \frac{\|\Delta g\|}{\sin \psi(\mathcal{Q}_b, \perp \mathcal{Q}_a)} \quad . \tag{48}$$

(2) Without any restrictions on f, Δg and $(\perp \mathcal{Q}_b) \cap \mathcal{Q}_a$,

$$\|f_k - f\| \leq \frac{\|\Delta g\| + \|Q_a Q_b f\| + (\|\Delta g\| + \|P_a f\|) \cdot \cos^k \psi(\mathcal{Q}_b, \perp \mathcal{Q}_a)}{1 - \cos \psi(\mathcal{Q}_b, \perp \mathcal{Q}_a)} \quad , \tag{49}$$

$k = 1 \to \infty$.

Proof. (1) Clearly, since $f \varepsilon \mathcal{Q}_b$ and $g = P_a f$, $g = Af$. Thus, $g + \Delta g = Af_o$ and $\Delta f = A^{-1}(\Delta g)$. If $\Delta g \varepsilon \mathcal{Q}_a$, $\|\Delta f\| \leq \mu \cdot \|\Delta g\|$ where

$$\mu = \sup_{x \varepsilon \mathcal{Q}_a} \frac{\|A^{-1}x\|}{\|x\|} \quad . \tag{50}$$

Let $x = Ay$. As we already know, under the assumption $(\perp \mathcal{Q}_b) \cap \mathcal{Q}_a = \{\phi\}$ $x \varepsilon \mathcal{Q}_a$ iff

$y \in \mathcal{\theta}_b$ and we can therefore write

$$\mu = \sup_{y \in \mathcal{\theta}_b} \frac{\|y\|}{\|y - Q_a P_b y\|} = \sup_{y \in \mathcal{\theta}_b} \frac{\|y\|}{\|P_a P_b y\|} = \frac{1}{\sin \psi(\mathcal{\theta}_b, \perp \mathcal{\theta}_a)} \quad , \tag{51}$$

establishing Eq. (48), Q.E.D.

(2) Let $g_1 = Af$. Then, $P_a g_1 = P_a f = g$ and

$$f_k - f = f_o - f + f_k - f_o = A^{-1}(g + \Delta g - g_1) - \sum_{r=k}^{\infty} (Q_a P_b)^r (g + \Delta g) \quad .$$

Or, since $g - g_1 = -Q_a g_1 = -Q_a(1 - Q_a P_b)f = -Q_a Q_b f$,

$$f_k - f = A^{-1}(\Delta g - Q_a Q_b f) - (Q_a P_b)^k A^{-1}(P_a f + \Delta g) \quad . \tag{52}$$

Combining Eq. (44) with $\|Q_a P_b\| = \cos \psi(\mathcal{\theta}_b, \perp \mathcal{\theta}_a)$ we immediately derive Eq. (49) by norming both sides of Eq. (52), Q.E.D.

Comment. The second and third terms in the numerator of Eq. (49) account, respectively, for aliasing (which invalidates the assumption $f \in \mathcal{\theta}_b$) and truncation error. Also observe that under the conditions prevailing in (1), the bound in Eq. (48) is sharper than that obtained by setting $Q_b f = \phi$ and letting $k \to \infty$ in Eq. (49).

In part (a_3) of theorem 1 it has been shown that the recursion Eq. (36) converges to f for every g in the range of the operator $A = 1 - Q_a P_b$ <u>even</u> if $\mu(\mathcal{\theta}_b, \perp \mathcal{\theta}_a) = 0$. The proof hinges on the uniqueness criterion

$$\mathcal{\theta}_b \cap (\perp \mathcal{\theta}_a) = \{\phi\} \tag{53}$$

of (a_1). As expected, if $g \notin R(A)$, $\lim \|f_k\| = \infty$.

<u>Corollary 2</u>. Assume that Eq. (53) is valid and the vector g is <u>not</u> in the range of A. Let $\{f_k\}$ denote the sequence of approximations generated by Eq. (36) under the initialization $f_1 = g$. Then,

$$\lim_{k \to \infty} \|f_k\| = \infty \quad . \tag{54}$$

<u>Proof</u>. If Eq. (54) doesn't hold, $\liminf \|f_k\| < \infty$ and there exists a subsequence $\{f_{k'}\}$ which is bounded in norm. Since Hilbert space is weakly compact[1] one can extract from this subsequence another subsequence $\{f_{k''}\}$ which converges weakly* to some

---

*$f_{k''} \rightharpoonup f_o$ if $\lim(f_{k''}, x) = (f_o, x)$ for every $x \in \mathcal{H}$.

element $f_o \varepsilon \mathcal{H}$; i.e., $f_{k''} \rightarrow f_o$. Now from Eq. (36),

$$(1 - Q_a P_b) f_k = \sum_{r=o}^{k-1} (1 - Q_a P_b)(Q_a P_b)^r g = g - (Q_a P_b)^k g \tag{55}$$

and using Eq. (53) and von Neumann's theorem on alternating projections,

$$\lim_{k\to\infty} (1 - Q_a P_b) f_k = g \quad . \tag{56}$$

However, because $1 - Q_a P_b$ is a bounded operator, $(1 - Q_a P_b) f_{k''} \rightarrow (1 - Q_a P_b) f_o$. But according to Eq. (56), $(1 - Q_a P_b) f_{k''} \rightarrow g$ and since the weak and strong limits of a sequence must coincide,[1] $(1 - Q_a P_b) f_o = g$. This means that $g \varepsilon R(A)$, a contradiction, Q.E.D.

To pin down the source of the blow-up in Eq. (54), suppose $Af_o = g_o$ and Eq. (36) is initialized with $f_1 = g_o + \Delta g$. Then, for $k \geq 1$

$$f_k = f_o - (Q_a P_b)^k f_o + \delta_{k-1} \tag{57}$$

where

$$\delta_{k-1} = \sum_{r=o}^{k-1} (Q_a P_b)^r \cdot \Delta g \quad . \tag{58}$$

If $g_o + \Delta g \notin R(A)$, corollary 2 yields $\lim \| \delta_{k-1} \| = \infty$ since $\lim (Q_a P_b)^k f_o = \phi$.* Now in the error expression

$$\varepsilon_k = f_o - f_k = (Q_a P_b)^k f_o - \delta_{k-1} \quad , \tag{59}$$

the norm of the first term is monotonically decreasing and it is evident that the algorithm should be terminated after a certain "optimal" number of steps. Unfortunately, the exact number depends both on the particular f_o under consideration and the growth rate of $\| \delta_{k-1} \|$ versus the decay rate of $\| (Q_a P_b)^k f_o \|$. Although the situation is not completely desperate it is definitely unsatisfactory but an in-depth discussion must be reserved for a future publication.

C. Discussion and an Example

It appears that in the majority of applications the modeling of image restoration problems is usually carried out with the aid of simplifying physical idealizations which

---

*Note that $\lim \| \delta_{k-1} \| = \infty$ for $\Delta g \notin R(A)$ irrespective of the size of $\| \Delta g \|$.

invariably endow the operator $P_a P_b$ with a set of nonzero eigenvalues possessing 0 as a limit point. All these problems are improperly posed.

Theorem 2. Let 0 be a limit point of the set of nonzero eigenvalues of the operator $P_a P_b$. Then, $\psi(\mathcal{M}_b, \perp \mathcal{M}_a) = 0$.

Proof. By hypothesis, there exists a sequence of nonzero scalars λ_i with limit 0 and a sequence of vectors $x_i \in \mathcal{H}$ such that

$$P_a P_b x_i = \lambda_i x_i \quad , \quad \|x_i\| = 1 \quad ; \quad i = 1 \to \infty \quad . \tag{60}$$

Clearly, $\lambda_i \neq 0$ implies $x_i \in \mathcal{M}_a$. Thus, $x_i = P_a x_i$ and the restriction of $P_a P_b$ to $\mathcal{M}_a$ coincides with the (positive) self-adjoint operator $P_a P_b P_a$. The sequence $\{x_i\}$ can therefore be chosen orthonormal[1] and as a consequence, the sequence $\{P_b x_i\}$ is orthogonal. Explicitly,

$$(P_b x_i, P_b x_j) = (P_b^2 x_i, x_j) = (P_b x_i, P_a x_j) = (P_a P_b x_i, x_j) = \lambda_i (x_i, x_j) = \lambda_i \delta_{ij} \quad , \tag{60a}$$

$i, j = 1 \to \infty$. In particular,

$$\lambda_i = \|P_b x_i\|^2 > 0 \quad , \quad i = 1 \to \infty \quad . \tag{61}$$

To calculate $\psi(\mathcal{M}_b, \perp \mathcal{M}_a)$ we employ the formula

$$\cos \psi(\mathcal{M}_b, \perp \mathcal{M}_a) = \sup_{x \in \mathcal{M}_b} \frac{\|Q_a x\|}{\|x\|} \tag{62}$$

given in Eq. (24) and substitute $P_b x_i (\in \mathcal{M}_b)$ for x. Using Eqs. (60) and (61),

$$\frac{\|Q_a P_b x_i\|}{\|P_b x_i\|} = \frac{\sqrt{\|P_b x_i\|^2 - \|P_a P_b x_i\|^2}}{\|P_b x_i\|} = \sqrt{1 - \lambda_i} \to 1 \tag{63}$$

and it follows that the sup on the right-hand side of Eq. (62) equals 1; i.e., $\psi(\mathcal{M}_b, \perp \mathcal{M}_a) = 0$, Q.E.D.

Comment 1. Norming both sides of Eq. (60) it is seen that $1 \geq \lambda_i$ with equality iff $x_i \in \mathcal{M}_b \cap \mathcal{M}_a$. Hence, $\mathcal{M}_b \cap \mathcal{M}_a = \{\phi\}$ implies $1 > \lambda_i > 0$, $i = 1 \to \infty$. Moreover, if $\mathcal{M}_b \cap (\perp \mathcal{M}_a) = \{\phi\}$, the completeness of the x_i's in $\mathcal{M}_a$ also entails that of the $P_b x_i$'s in $\mathcal{M}_b$. Taking the orthogonality of the sequence $\{P_b x_i\}$ into account, it suffices for the proof to show that $h = \phi$ is the only vector in $\mathcal{M}_b$ satisfying $(h, P_b x_i) = 0$ for all i. For such an h, $(P_a h, x_i) = 0$, $i = 1 \to \infty$ and from the completeness of the x_i's in $\mathcal{M}_a$, $P_a h = \phi$. Thus, $h \in \mathcal{M}_b \cap (\perp \mathcal{M}_a) = \{\phi\}$ and $h = \phi$, Q.E.D. The reader has now undoubtedly perceived that under the assumptions

$$\mathcal{Q}_b \cap \mathcal{Q}_a = \{\phi\} = \mathcal{Q}_b \cap (\perp \mathcal{Q}_a) \quad , \tag{64}$$

the x_i's enjoy all the remarkable properties of the prolate spheroidal wavefunctions.[4] These properties simply reflect the structure of $P_a P_b$ as the composition of two orthogonal projection operators.

Example. Consider the Hilbert space $\mathcal{H}$ of all functions f(t) square-integrable over $-\infty < t < \infty$ equipped with the inner product*

$$(f, g) = \int_{-\infty}^{\infty} f(t)\, g^*(t)\, dt \quad . \tag{65}$$

To indicate that f(t) and F(ω) are Fourier transform pairs we write $f(t) \leftrightarrow F(\omega)$. Of course,

$$F(\omega) = \int_{-\infty}^{\infty} f(t) e^{-i\omega t}\, dt \tag{66}$$

is the L_2-Fourier transform of $f(t) \in \mathcal{H}$.

For $a > 0$ prescribed and finite, let $\mathcal{Q}_a$ denote the subset of composed of all functions which vanish (a.e.) in $|t| > a$. Similarly, for a prescribed finite $b > 0$ let $\mathcal{Q}_b$ contain all functions whose Fourier transforms vanish (a.e.) in $|\omega| > b$. It is easily shown that 1) $\mathcal{Q}_a$ and $\mathcal{Q}_b$ are CLMS, 2) that $\perp \mathcal{Q}_a$ is the set of all functions vanishing (a.e.) in $|t| \leq a$ and 3) that $\perp \mathcal{Q}_b$ is the set of all functions whose Fourier transforms vanish (a.e.) in $|\omega| \leq b$. Furthermore, letting

$$g_a(t) = \begin{cases} 1, & |t| \leq a \\ 0, & |t| > a \end{cases} \tag{67}$$

and

$$g_b(\omega) = \begin{cases} 1, & |\omega| \leq b \\ 0, & |\omega| > b \end{cases} ,$$

it is evident that if $f(t) \leftrightarrow F(\omega)$, $P_a f = g_a(t) f(t)$ and $P_b f \leftrightarrow g_b(\omega) F(\omega)$. Explicitly,

$$P_b f = \frac{1}{2\pi} \int_{-b}^{b} F(\omega)\, e^{i\omega t}\, d\omega = \int_{-\infty}^{\infty} \frac{\sin b(t-\tau)}{\pi(t-\tau)} f(\tau)\, d\tau \quad . \tag{69}$$

Thus, $\mathcal{Q}_b$ is the set of all functions bandlimited to the (radian) frequency interval $|\omega| \leq b$ and $\mathcal{Q}_a$ the set of all functions time-limited to the interval $|t| \leq a$. The

---

The function g(t) is the complex conjugate of g(t).

corresponding image restoration problem can be stated as follows: A function f(t) is known to be bandlimited to the interval $-b \leq \omega \leq b$ but only its segment over $-a \leq t \leq a$ is available. Reconstruct f(t) for $|t| > a$. In our formulation, $f \in \mathcal{B}_b$ and the projection $g = P_a f = g_a(t) f(t)$ onto $\mathcal{B}_a$ is given. By theorem 1, the reconstruction of f(t) can be accomplished with the aid of the recursion Eq. (36) provided $\mathcal{B}_b \cap (\perp \mathcal{B}_a) = \{\phi\}$. But this is obvious since any $f(t) \in \mathcal{B}_b$ is automatically an entire function of the complex variable t and can vanish in $|t| \leq a$ iff it is the zero function. (The same reasoning also yields $\mathcal{B}_b \cap \mathcal{B}_a = \{\phi\} = (\perp \mathcal{B}_b) \cap \mathcal{B}_a$.) Hence, translating Eq. (36) into words, it is seen that the sequence $\{f_k(t)\}$ generated by the 2-rule program given below converges to f(t) in norm:*

(R_1) Set $f_1(t) = g_a(t) f(t)$, the initial datum.

(R_2) For $k \geq 1$ bandlimit $f_k(t)$ to the interval $|\omega| \leq b$ (use a low-pass filter or FFT) and correct the resultant waveform over $|t| \leq a$ to agree with $f_1(t)$. Call the final function $f_{k+1}(t)$ and repeat (R_2).

This 2-step procedure which emerges as a special case of theorem 1 is precisely the new algorithm published recently by Papoulis[5] who based the proof of convergence on the detailed properties of the prolate spheroidal wavefunctions. The latter are eigenfunctions of the operator $P_a P_b$. In fact, for $x_i(t) \in \mathcal{B}_a$,

$$P_b x_i = \int_{-a}^{a} \frac{\sin b(t-\tau)}{\pi(t-\tau)} x_i(\tau)\, d\tau \tag{70}$$

and $P_a P_b x_i = \lambda_i x_i$ is equivalent to

$$\int_{-a}^{a} \frac{\sin b(t-\tau)}{\pi(t-\tau)} x_i(\tau)\, d\tau = \lambda_i x_i(t) \quad , \quad -a \leq t \leq a \quad , \tag{71}$$

a familiar integral equation.[4] As is well known,[4] $1 > \lambda_i > 0$ for $i = 1 \to \infty$, $\lambda_i \to 0$ and the x_i's are complete in $\mathcal{B}_a$. Since $\mathcal{B}_a \cap (\perp \mathcal{B}_a) = \{\phi\}$, the $P_b x_i$'s are complete in $\mathcal{B}_b$. In addition, invoking theorem 2, $\psi(\mathcal{B}_b, \perp \mathcal{B}_a) = 0$** which means that the reconstruction problem is improperly posed, a conclusion also reached by Viano[7] in a different way.

---

*As is easily shown, it also converges to f(t) uniformly over $-\infty < t < \infty$ (Ref. 5).

**But interestingly enough, by exploiting the completeness of the $P_b x_i$'s in $\mathcal{B}_b$ and using Eq. (60), the formula

$$\cos \psi(\mathcal{B}_b, \mathcal{B}_a) = \sup_{x \in \mathcal{B}_b} \frac{\| P_a P_b x \|}{\| x \|} \tag{71a}$$

yields $\psi(\mathcal{B}_b, \mathcal{B}_a) = \cos^{-1}\sqrt{\lambda_1}$. Since $1 > \lambda_1 > 0$ we have $0 < \psi(\mathcal{B}_b, \mathcal{B}_a) < \frac{\pi}{2}$, a result due to Landau and Pollack.[6]

The author wishes to take this opportunity to record his indebtedness to Refs. 5 and 6 and to thank his colleagues Profs. Laemmel and Papoulis for engaging in several interesting discussions regarding the subject matter of this report.

Joint Services Technical Advisory Committee
F44620-74-C-0056

D.C. Youla

National Science Foundation
ENG 74-13054

REFERENCES

1. A.C. Zaanen, "Linear Analysis," Interscience Publishers, Inc. (1953).
2. T. Kato, "Perturbation Theory for Linear Operators," (New York: Springer-Verlag, 1966).
3. J. von Neumann, "The Geometry of Orthogonal Spaces," Princeton University Press, II (1950).
4. D. Slepian and H.O. Pollack, "Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty - I," Bell System Tech. J., 40, 43-64 (January 1961).
5. A. Papoulis, "A New Algorithm in Spectral Analysis and Band-Limited Extrapolation," IEEE Trans. on Circuits and Systems, CAS-22, No. 9 (September 1975).
6. H.J. Landau and H.O. Pollack, "Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty - II," Bell System Tech. J., 40, 65-84 (January 1961).
7. G.A. Viano, "On the Extrapolation of Optical Image Data," J. Math. Phys., 17, No. 7, 1160-1165 (July 1976).

AN APPROACH TO THE PATTERN CLASSIFICATION AND ESTIMATION PROBLEM USING THE CONCEPT OF STOCHASTIC DIFFERENTIAL EQUATIONS

L. Kurz

Grenander[1] has considered one approach to the problem of employing statistical methods to pattern classification. His development, which makes use of his earlier work in statistical inference,[2] can be divided into three somewhat arbitrary sections. First, one selects a method for describing the pure or undisturbed images so that a mathematical structure is available. Second, the deformation or noise disturbances of the pure images are characterized. Third, a selection scheme or decision rule for distinguishing between images is chosen. These three sub-problems are analogous to the problems of alphabet description, channel characterization and receiver structure selection which arise in communication systems. The description of the pure images and the means for their generation as well as certain transformation properties are accomplished by setting up an appropriate algebraic structure. The probabilistic deformation grammar is introduced so that appropriate measures can be used to describe the disturbances and provide a background in which the methods of statistical inference can be applied to pattern classification problems. Because the goals of the development are broad, the choice of a decision rule is arrived at from an intuitive approach of distance and maximum likelihood. As such, the well known Random Nikodym derivative (RND) or generalized likelihood ratio is introduced and serves as the basis of the *decision rules*.

The RND may be used successfully in classification and maximum likelihood estimation (MLE) problems. In particular, a method of finding the appropriate RND's of a one-dimensional Gaussian process, which is equivalent to the Wiener process or one of its variants, is embedded in an n-dimensional Markov process which is the solution of a vector stochastic differential equation. Through this embedding procedure, one makes use of the structure of vector Markov processes, their equivalence conditions and the corresponding RND. Though the procedure applies to classification and estimation problems, specific examples will be given for MLR only. The extension to classification problems is routine. Details are given for a system transfer function described by an all-pole model. Other models may be treated in a similar manner.

A. Markov Processes and Stochastic Differential Equations

The stochastic differential equations to be considered are of the form[3]

$$d\xi(t) = a(t, \xi(t))dt + \sum_{k=1}^{\ell} b_k(t, \xi(t))\, dW_k(t) \qquad (1)$$

where $\xi(t)$ takes values in R^n, $W_k(t)$ are Wiener processes that are mutually independent

and $a(t, \xi(t))$ and $b_k(t, \xi(t))$ are n-dimensional vector coefficients. If the initial condition is specified, the solution to Eq. (1) is the process with sample functions $\xi(t)$ which are given by

$$\xi(t) = \xi(0) + \int_0^t a(s, \xi(s))ds + \sum_{k=1}^{\ell} \int_0^t b_k(s, \xi(s))\, dW_k(s) \quad , \quad t \geq 0 \tag{2}$$

The question of existence and uniqueness of solutions to Eq. (1) are answered by the following theorems.

Theorem 1: Suppose that the following conditions are fulfilled:

1. $\xi(0)$ does not depend on the processes $W_1(t), \cdots, W_\ell(t)$,

2. For every $c > 0$ there exists an L_c such that for $|x| \leq c$, $|y| \leq c$, the inequality

$$T|a(t,x) - a(t,y)|^2 + \sum_{k=1}^{\ell} |b_k(t,x) - b_k(t,y)|^2 \leq L_c |x - y|^2 \tag{3}$$

is fulfilled,

3. There exists a K such that

$$|a(t,x)|^2 + \sum_{k=1}^{\ell} |b_k(t,x)|^2 \leq K(1 + |x|^2) \tag{4}$$

4. $a(t,x)$, $b_1(t,x)$, $\cdots$, $b_\ell(t,x)$ are continuous with respect to the set of variables $x \in R^n$, $t \in [0, T]$, then Eq. (1) has a continuous and bounded solution with probability one. Further, the solution is unique in the sense that any two such solutions coincide with probability one at all points.

Theorem 2: If the conditions of Theorem 1 are satisfied, the solution to Eq. (1) is a Markov process.

Suppose that $\xi_1(t)$ and $\xi_2(t)$ are solutions of the stochastic differential equations

$$\xi_i(t) = \xi_i(0) + \int_0^t a_i(s, \xi_i(s))ds + \sum_{k=1}^{\ell} \int_0^t b_k(s, \xi_i(s))\, dW_k(s) \quad , \quad i = 1, 2 \tag{5}$$

the coefficients of which satisfy the conditions of Theorem 1. The following theorem establishes sufficient conditions for the equivalence of the measures μ_1 and μ_2 corresponding to the processes $\xi_1(t)$ and $\xi_2(t)$, and also a general expression for the associated RND.

Theorem 3: Suppose that the following conditions are satisfied:

1. There exists functions $a_1(t,x), \cdots, a_\ell(t,x)$ such that

$$a_2(t,x) - a_1(t,x) = \sum_{k=1}^{\ell} a_k(t,x)\, b_k(t,x) \tag{6}$$

2. The functions $a_1(t,x), \cdots, a_\ell(t,x)$ and $a(t,x)$ given by

$$a(t,x) = -\frac{1}{2}\sum_{j=1}^{\ell} a_j^2(t,x) \tag{7}$$

are continuous over the set of variables,

3. The distributions of $\xi_1(0)$ and $\xi_2(0)$ are absolutely continuous with respect to one another and the density of the distribution of $\xi_2(0)$ with respect to the distribution of $\xi_1(0)$ is $p_0(x)$, then the measures μ_1 and μ_2 will be equivalent and

$$\log \frac{d\mu_2}{d\mu_1}(\xi_1(t) = \log p_0(\xi_1(0)) + \int_0^T a(t, \xi_1(t))dt + \sum_{j=0}^{\ell} \int_0^T a_j(t, \xi_1(t))\, d W_j(t) \tag{8}$$

Equation (8) is a general expression for the log of the RND of two equivalent n-dimensional Markov processes. It will be necessary to use this equation to derive an expression for the RND corresponding to two equivalent one-dimensional processes which are components of $\xi_1(t)$ and $\xi_2(t)$. This is facilitated by the following theorem.

Theorem 4: Suppose that $\xi_1(t)$ and $\xi_2(t)$ are processes defined on $[0,T]$ and taking values in R^n. Denote by $\Phi_n[0,T]$ the set of all functions $x(t)$ that are defined on $[0,T]$ and that take values in R^n. Denote by $F_n[0,T]$ the minimal σ-algebra of subsets $\Phi_n[0,T]$ that contains all cylindrical sets. Let $\mathcal{J}$ be a $F_n[0,T] \times F_{n'}[0,T]$ measurable mapping of $\Phi_n[0,T]$ into $\Phi_{n'}[0,T]$. Suppose that $\eta_1(t) = \mathcal{J}\xi_1(t)$, $\eta_2(t) = \mathcal{J}\xi_2(t)$, the μ_i are measures corresponding to the processes $\xi_i(t)$ on $F_n[0,T]$ and that ν_i are measures corresponding to the processes $\eta_i(t)$ on $F_{n'}[0,T]$, then if μ_2 is equivalent to μ_1, it follows that ν_2 is equivalent to ν_1 and

$$\frac{d\nu_2}{d\nu_1}(\eta_1(t)) = E\Big[\frac{d\mu_2}{d\mu_1}(\xi_1(t))/\eta_1(t) \quad , \quad t \in [0,T]\Big] \tag{9}$$

B. Solution of Estimation Problems

The general vector Markov process is specified by Eq. (2) which in turn is completely determined by the coefficients $a(t, \xi(t))$, $b_k(t, \xi(t))$ and the initial condition $\xi(0)$. For Gaussian processes certain significant simplifications occur. If $\xi(t)$ is Gaussian, then the components of $a(t, \xi(t))$ can only be linear functions of the components of $\xi(t)$

and $b(t, \xi(t))$ cannot depend on $\xi(t)$. With this in mind, it is of interest to represent the one-dimensional reference processes (i.e., the Wiener process and its variants) as components of vector Markov processes.

Consider the free Wiener process defined by

$$\underset{\sim}{W}_0(t) = \eta + \int_0^t d\,W(t) \tag{10}$$

where η is an $N(0, 1)$ random variable that is independent of the Wiener process $W(t)$. Obviously $\{\underset{\sim}{W}_0(t)\}$ is a one-dimensional Markov process with

$$a(t, \underset{\sim}{W}_0(t)) = 0 \tag{11}$$

$$b(t, \underset{\sim}{W}_0(t)) = 1 \tag{12}$$

and

$$\underset{\sim}{W}_0(0) = \eta \tag{13}$$

Similarly, the first order free Wiener process is defined by

$$\underset{\sim}{W}_1(t) = \eta_0 + \eta_1 t + \int_0^t (t - u)\, d\,W(u) \tag{14}$$

where η_0 and η_1 are independent $N(0, 1)$ random variables.

By formally differentiating Eq. (14) we have

$$\underset{\sim}{W}_1^{(1)}(t) = \eta_1 + \int_0^t d\,W(u) \tag{15}$$

If we denote by $\xi(t)$ the following 2-dimensional process

$$\xi(t) = \begin{bmatrix} \underset{\sim}{W}_1^{(1)}(t) \\ \\ \underset{\sim}{W}_1(t) \end{bmatrix} = \begin{bmatrix} \eta_1 \\ \\ \eta_0 \end{bmatrix} + \int_0^t \begin{bmatrix} 0 \\ \\ \underset{\sim}{W}_1^{(1)}(u) \end{bmatrix} du + \int_0^t \begin{bmatrix} 1 \\ \\ 0 \end{bmatrix} d\,W(u) \tag{16}$$

then it is a Markov process. For the general n-th-order process we define the $(n+1)$-dimensional Markov process

$$\xi(t) = \begin{bmatrix} \underset{\sim}{W}_n^{(n)}(t) \\ \underset{\sim}{W}_n^{(n-1)}(t) \\ \vdots \\ \underset{\sim}{W}_n^{(1)}(t) \\ \underset{\sim}{W}_n(t) \end{bmatrix} \tag{17}$$

which satisfies the vector equation having coefficients

$$a(t, \xi(t)) = \begin{bmatrix} 0 \\ \underset{\sim}{W}_n^{(n)}(t) \\ \underset{\sim}{W}_n^{(n-1)}(t) \\ \vdots \\ \underset{\sim}{W}_n^{(2)}(t) \\ \underset{\sim}{W}_n^{(1)}(t) \end{bmatrix}; \quad b(t, \xi(t)) = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} \tag{18}$$

and the initial condition

$$\xi(0) = \begin{bmatrix} \eta_n \\ \eta_{n-1} \\ \vdots \\ \eta_1 \\ \eta_0 \end{bmatrix} \tag{19}$$

where the η_i's are independent $N(0, 1)$ random variables and the $(n+1)^{st}$-component of Eq. (17) is the original one-dimensional process. It is important to note that although the formal differentiation of Eq. (10) is valid, the correct interpretation of the differential of a Markov process requires the use of Ito's formula.[4] Equations (10) through (19) characterize the free Wiener process $\{\underset{\sim}{W}_n(t)\}$. By simply changing the initial conditions associated with these equations, we obtain the equations for the integrated Wiener process $\{W_n(t)\}$. The desired equations have the coefficients

$$a(t, \xi(t)) = \begin{bmatrix} 0 \\ W_n^{(n)}(t) \\ W_n^{(n-1)}(t) \\ \vdots \\ W_n^{(1)}(t) \end{bmatrix} ; \quad b(t, \xi(t)) = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} \tag{20}$$

and the initial condition $\xi(0) = 0$ with probability one. Equations (10) through (20) are of a simple form and describe processes that are generated by a "driving process" which is a one-dimensional Wiener process. By choosing different coefficients than those of Eq. (18), we obtain different processes.

C. Likelihood Ratios (RNDs) For Estimation of Parameters of Lumped Systems Without Numerator Dynamics

Consider the process $\{x(t)\}$ having power spectrum of the form

$$S(\omega) = \frac{1}{(\omega^2 + S_1^2)(\omega^2 + S_2^2) \cdots (\omega^2 + S_n^2)} \tag{21}$$

It is known that the process $\{x(t)\}$ is $(n-1)$-times differentiable and equivalent to $\underset{\sim}{W}_{n-1}(t)$. Although $\{x(t)\}$ is not a one-dimensional Markov process, the n-dimensional process consisting of x(t) and its (n-1) derivatives is Markov. To find the appropriate vector equation, consider the vector process

$$\xi_2(t) = \begin{bmatrix} x^{(n-1)}(t) \\ x^{(n-2)}(t) \\ \vdots \\ x(t) \end{bmatrix} \tag{22}$$

The coefficients $a_2(t, \xi_2(t))$ and $b_1(t, \xi_2(t))$ can be found from the infinitesimal properties of the process $\xi_2(t)$ using the formulas

$$a_2(t, \xi_2(t)) = \begin{bmatrix} a_{21}(t, \xi_2(t)) \\ \vdots \\ a_{2n}(t, \xi_2(t)) \end{bmatrix} = \lim_{h \downarrow 0} \left[\frac{\xi_2(t+h) - \xi_2(t)}{h} \Big/ \xi_2(t) \right] \tag{23}$$

$$b_1(t, \xi_2(t)) = \begin{bmatrix} b_{11}(t, \xi_2(t)) \\ \vdots \\ b_{1n}(t, \xi_2(t)) \end{bmatrix} = \lim_{h \downarrow 0} E\left[\frac{(\xi_2(t+h) - \xi_2(t))^2}{h} \Big/ \xi_2(t) \right] \tag{24}$$

The conditional expectations of Eqs. (22) and (23) can be evaluated because the variables $x(t)$, $x^{(1)}(t)$, $\cdots$, $x^{(n-1)}(t)$ are Gaussian and their n-dimensional covariance matrix can be determined using Equation (21). Alternatively, these coefficients may be determined from the defining equations for $x^{(i)}(t)$. For all i such that $0 \leq i \leq n-2$, the sample functions $x^{(i)}(t)$ are differentiable a.s. so that the defining equations yield the coefficients

$$a_{2j} = x^{(n-j+1)}(t) = \frac{dx^{(n-j)}(t)}{dt}$$

and (25)

$$b_{1j} = 0 \qquad 1 < j \leq n$$

To determine a_{21} and b_{11}, we use the following recursion relation

$$y_k(t) = S_{k+1}\, y_{k+1}(t) + y_{k+1}^{(1)}(t) \qquad 1 \leq k \leq n-2 \tag{26}$$

with the end conditions

$$y_{n-1}(t) = S_n\, x(t) + x^{(1)}(t) \tag{27}$$

and

$$d\,W(t) = y_0(t)\, dt = S_1 y_1(t)\, dt + dy_1(t) \tag{28}$$

where $W(t)$ is the Wiener process. Equation (28) is interpreted in the Ito sense.[4] The correct results can also be obtained through the formal procedure of rewriting Eq. (26) as

$$n(t) = y_0(t) = \frac{d\,W(t)}{dt} = S_1 y_1(t) + \frac{dy_1(t)}{dt} \tag{29}$$

where $n(t)$ is the white noise process. Using these equations, we find that

$$a_2(t, \xi_2(t)) = \begin{bmatrix} m^{(n)}(t) - \sum_{i=1}^{n} (x^{(n-i)}(t) - m^{(n-i)}(t))\, S_n(i) \\ x^{(n-1)}(t) \\ x^{(n-2)}(t) \\ \vdots \\ x^{(1)}(t) \end{bmatrix}; \quad b_1(t, \xi_2(t)) = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{30}$$

where m(t) is the mean function of the observed process x(t) and

$S_n(j)$ = sum of all different products of the parameters $s_1, s_2, \cdots, s_n$ taken j at a time.

For example, if n = 3, $S_3(1) = s_1 + s_2 + s_3$, $S_3(2) = s_1s_2 + s_1s_3 + s_2s_3$ and $S_3(3) = s_1s_2s_3$.

Now that the coefficients of Eq. (30) have been derived, the next task is to choose an appropriate n-dimensional reference process, say $\xi_1(t)$, which can be used to find the desired likelihood equation. Motivated by the facts that the observed process x(t) is equivalent to $W_{n-1}(t)$ and that the b-coefficient matrix of $\xi_1(t)$ must be equal to the b-matrix of Eq. (30), it is natural to select $\xi_1(t)$ to be

$$\xi_1(t) = \begin{bmatrix} \underset{\sim}{W}_{n-1}^{(n-1)}(t) \\ \underset{\sim}{W}_{(n-1)}^{(n-2)}(t) \\ \vdots \\ \underset{\sim}{W}_{(n-1)}(t) \end{bmatrix} \tag{31}$$

This process was discussed in Section B; its coefficient matrices are given by Equation (18). Using the above equations, it is a straightforward matter to verify the conditions of Theorem 3 and show that the measures P_1 and P_2 of the two n-dimensional processes $\xi_1(t)$ and $\xi_2(t)$, respectively, are equivalent. If the observation interval is [0, T], then the logarithm of the generalized likelihood ratio is given by Eq. (8) which reduces to

$$\log \frac{dP_2}{dP_1}(\xi_1(t)) = \log p_0(\xi_1(0)) + \int_0^T a(t, \xi_1(t))dt + \int_0^T a_1(t, \xi_1(t))\, d\,W(t) \tag{32}$$

where $a_1(t, \xi_1(t))$ satisfies the equation

$$a_1(t, \xi_1(t)) \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = a_2(t, \xi_1(t)) - a_1(t, \xi_1(t)) = \begin{bmatrix} m^{(n)}(t) - \sum_{i=1}^{n} [\underset{\sim}{W}_{n-1}^{(n-i)}(t) - m^{(n-i)}(t)] S_n(i) \\ \underset{\sim}{W}_{n-1}^{(n-1)}(t) \\ \vdots \\ \underset{\sim}{W}_{n-1}^{(1)}(t) \end{bmatrix} -$$

$$\begin{bmatrix} 0 \\ \underset{\sim}{W}_{n-1}^{(n-1)}(t) \\ \vdots \\ \underset{\sim}{W}_{n-1}^{(1)}(t) \end{bmatrix} \tag{33}$$

or

$$a_1(t, \xi_1(t)) = m^{(n)}(t) - \sum_{i=1}^{n} (\underset{\sim}{W}_{n-1}^{(n-i)}(t) - m^{(n-i)}(t))\, S_n(i) \tag{34}$$

and

$$a(t, \xi_1(t)) = \frac{1}{2}\, a_1^2(t, \xi_1(t)) \tag{35}$$

To obtain a useful form of Eq. (32), which can easily be implemented by a practical estimator, it is necessary to evaluate the last term which is a stochastic or Ito integral. Although this is, in general, a difficult task, it can be readily accomplished in the present case. Because Eq. (34) contains terms which depend linearly on the derivatives of $\underset{\sim}{W}_{n-1}(t)$, the integral can be evaluated by using Ito's formula and the smoothness properties of the individual terms; the result is

$$\begin{aligned} G[W_{n-1}(t)] \equiv \int_0^T a_1(t, \xi_1(t))\, dW(t) &= m^{(n)}(T)\, \underset{\sim}{W}_{n-1}^{(n-1)}(T) - m^{(n)}(0)\, \underset{\sim}{W}_{n-1}^{(n-1)}(0) + \frac{1}{2}\, S_n(1) \\ &\times \left[\underset{\sim}{W}_{n-1}^{2(n-1)}(T) - \underset{\sim}{W}^{2(n-1)}(0) + T\right] - \sum_{j=1}^{n} S_n(j) \left\{ \underset{\sim}{W}_{n-1}^{(n-1)}(T) \left[\underset{\sim}{W}_{n-1}^{(n-j)}(T) \right.\right. \\ &\left.\left. - m^{(n-j)}(T)\right] - \underset{\sim}{W}_{n-1}^{(n-1)}(0) \left[\underset{\sim}{W}_{n-1}^{(n-j)}(0) - m^{(n-j)}(0)\right] \right\} + \sum_{j=2}^{n} S_n(j) \int_0^T \\ &\underset{\sim}{W}_{n-1}^{(n-1)}(t)\, \underset{\sim}{W}_{n-1}^{(n-j+1)}(t)\, dt - \int_0^T \underset{\sim}{W}_{n-1}^{(n-1)}(t) \left[\sum_{j=0}^{n} S_n(j)\, m^{(n+1-j)}(t)\right] dt \end{aligned} \tag{36}$$

where

$$S_n(0) \equiv 1 \quad \text{for all } n$$

Equation (33) represents the likelihood ratio of the measures of the two n-dimensional processes $\xi_1(t)$ and $\xi_2(t)$. However, because we can only observe the output process x(t) we are really interested in obtaining the likelihood ratio of the one-dimensional processes P_x with respect to $P_{\underset{\sim}{W}_{n-1}}$. This is accomplished by using Theorem 4 which yields the result

$$\frac{dP_x}{dP_{\underset{\sim}{W}_{n-1}}}(\underset{\sim}{W}_{n-1}(t)) = E\left\{\frac{dP_2}{dP_1}(\xi_1(t))/\underset{\sim}{W}_{n-1}(t),\ 0 \le 1 < T\right\} \tag{37}$$

To evaluate the conditional expectation of Eq. (37), we need only observe that if $\underset{\sim}{W}_{n-1}(t)$, $0 \le t \le T$, is known, then all the remaining components of $\xi_1(t)$, $0 \le t \le T$, can be

determined by differentiation. In other words, when conditioned on $\underset{\sim}{W}_{n-1}(t)$, $0 \leq t \leq T$, dP_2/dP_1 $(\xi_1(t)$ is a constant, so that

$$\frac{dP_x}{dP_{\underset{\sim}{W}_{n-1}}}(\underset{\sim}{W}_{n-1}(t)) = \frac{dP_2}{dP_1}(\xi_1(t)) \tag{38}$$

Finally, using the facts that $P_x \sim P_{\underset{\sim}{W}_{n-1}}$ and $P_2 \sim P_1$, where $\sim$ denotes equivalence, we can replace $\underset{\sim}{W}_{n-1}(t)$ by $x(t)$ in the above equations to obtain the desired result

$$\Lambda = \log \frac{dP_x}{dP_{\underset{\sim}{W}_{n-1}}}(x(t)) = \log p_0(x^{(n-1)}(0), \cdots, x(0) + \int_0^T a(t, x^{(n-1)}(t) \cdots x(t)\, dt + G[x(t)] \tag{39}$$

where $G[x(t)]$ is defined by Eq. (36), $a(t, x^{(n-1)}(t) \cdots x(t))$ is given by Eq. (36) and p_0 as in Eq. (38) is the ratio of the two n-dimensional Gaussian density functions.

D. Example of Application: MLE of Mean Function Parameter for a Two Pole Case

It can be shown that for a two pole system, Eq. (39) reduces to

$$\begin{aligned}\Lambda = \frac{1}{2}\Big\{&\log 4 s_1 s_2 (s_1 + s_2)^2 + [1 - (s_1 + s_2)]^2 x^{(1)}(0) + [1 - s_1 s_2 (s_1 + s_2)] x^2(0) - \\ &- (s_1 + s_2) x^{(1)}(T) - s_1 s_2 (s_1 + s_2) x^2(T) - 2 s_1 s_2 [x(T) x^{(1)}(T) - x(0) x^{(1)}(0)] + \\ &+ (s_1 + s_2) T + 2[2(s_1 + s_2) m^{(1)}(0) - f(0)] x^{(1)}(0) + 4 s_1 s_2 (s_1 + s_2) m(0) x(0) \\ &+ 2 f(T) x^{(1)}(T) - 2(s_1 + s_2)[m^{(1)}(0) + s_1 s_2 m^2(0)] - \int_0^T f^2(t)\, dt \\ &+ 2 s_1 s_2 \int_0^T f(t) x(t)\, dt + 2 \int_0^T [(s_1 + s_2) f(t) - f^{(1)}(t)] x^{(1)}(t)\, dt \\ &- (s_1^2 + s_2^2) \int_0^T x^{2(1)}(t)\, dt - s_1^2 s_2^2 \int_0^T x^2(t)\, dt \Big\}\end{aligned} \tag{40}$$

Consider the problem in which s_1 and s_2 are known and it is desired to estimate an unknown parameter, θ, of the mean function. This corresponds to the situation in which there is enough information available to specify the pole locations in the model used for the noise process. The transmitted signal, or pure image, however, cannot becompletely specified, but is of known form

$$m(t) = m(t; \theta) \tag{41}$$

It follows that f(t) depends on θ and is given by

$$f(t) = f(t;\theta) = m^{(2)}(t;\theta) + (s_1 + s_2)\, m^{(1)}(t;\theta) + s_1 s_2\, m(t;\theta) \tag{42}$$

Maximization of Λ with respect to θ results in the following likelihood equation

$$\begin{aligned} 0 = \frac{\partial}{\partial\theta}\Big\{ & [2(s_1+s_2)\, m^{(1)}(0;\theta) - f(0;\theta)]\, x^{(1)}(0) + 2 s_1 s_2 (s_1+s_2)\, m(0;\theta)\, x(0) \\ & + f(T;\theta)\, x^{(1)}(T) - (s_1+s_2)[m^{2(1)}(0;\theta) + s_1 s_2\, m^2(0;\theta)] - \tfrac{1}{2}\int_0^T f^2(t,\theta)\, dt \\ & + s_1 s_2 \int_0^T f(t;\theta)\, x(t)\, dt + \int_0^T [(s_1+s_2) f(t;\theta) - f^{(1)}(t;\theta)]\, x^{(1)}(t)\, dt \Big\} \end{aligned} \tag{43}$$

If $m(t;\theta)$ is of the form

$$m(t;\theta) = \theta\, g(t) \tag{44}$$

where g(t) is a known function of time and

$$f(t;\theta) = \theta\, G(t) \tag{45}$$

where

$$G(t) = g^{(2)}(t) + (s_1 + s_2)\, g^{(1)}(t) + s_1 s_2\, g(t) \tag{46}$$

Under these conditions, solution of *Eq. (43) yields the* following unbiased and efficient MLE for θ

$$\begin{aligned} \hat{\theta} = \Big[& [2(s_1+s_2)\, g^{(1)}(0) - G(0)]\, x^{(1)}(0) + 2 s_1 s_2 (s_1+s_2)\, g(0)\, x(0) + G(T)\, x^{(1)}(T) + \\ & + s_1 s_2 \int_0^T G(t)\, x(t)\, dt - \int_0^T [G^{(1)}(t) - (s_1+s_2) G(t)] \\ & \times \Big[2(s_1+s_2)[(g^{(1)}(0))^2 + s_1 s_2\, g^2(0)] + \int_0^T G^2(t)\, dt \Big]^{-1} \end{aligned} \tag{47}$$

This estimator depends on the observed waveform and its derivative and can be implemented.

Joint Services Technical Advisory Committee
F44620-74-C-0056

L. Kurz

REFERENCES

1. U. Grenander, "Foundations of Pattern Analysis," Quart. Appl. Math., 27, No. 1, 1-55 (April 1969).
2. U. Grenander, "Stochastic Processes and Statistical Inference," Arkiv for Matematik, 17, 577-591 (1950).
3. A.V. Skorokhod, "Studies in the Theory of Random Processes," (Massachusetts: Addison-Wesley, 1965).
4. K. Ito, "On a Formula Concerning Stochastic Differentials," Nagoya Math. J., 3 (1951).

A ROBUSTIZED KIEFER-WOLFOWITZ PROCEDURE WITH APPLICATION TO THRESHOLD ESTIMATION FOR BLOCK DATA PROCESSING

P. Kersten and L. Kurz

The stochastic approximation procedures of both Robbins-Monro,[7] and Kiefer-Wolfowitz[5] provide a class of recursive estimators which are intrinsically nonparametric and amenable to digital computation. The convergence properties of both of these procedures has been considered by many authors including Dvoretzky,[3] Burkholder,[2] and Sacks.[8] Dvoretzky considered the a.s. and mean-square convergences of these algorithms and Sacks concentrated on the asymptotic normality of these procedures.

Two modifications, which enhance the rate of convergence and the robustness of the Kiefer-Wolfowitz (KW) procedure, is the main thrust of this report. The first employs adaptive gain coefficients to ensure an asymptotically optimum rate of convergence and the second exploits non-linear pre-processing of the data to immunize the procedure against Burst Noise. Since many real-world forms of noise have distribution functions which may be modelled as a mixture distribution $F(x) = (1-\varepsilon)G(x) + \varepsilon H(x)$ where $H(x)$ is a high variance CDF, the latter problem is of considerable interest in communication, radar, sonar, picture processing, etc.

An application of the robustized KW and Robbins-Monro (RM) procedures to the problem of threshold estimation for block data processing is also presented.

A. A Robustized Kiefer-Wolfowitz Procedure

The KW procedure requires two positive null sequences and two samples per iteration to converge to the maximum (minimum) of the regression function. The first of the null sequences is $\{a_n(X_1,\cdots,X_n)\}$ where $a_n(\cdot) = A(X_1,\cdots,X_n)/n$ and $\Sigma a_n(\cdot) = \infty$ for every sequence $X_1,\cdots,X_n,\cdots$, and $A(X_1,\cdots,X_n)$ is a suitably selected nonparametric mean-square consistent estimator of some constant A. The second positive null sequence $\{c_n\}$ is chosen s.t. $\Sigma n^{-2} c_n^{-2} A_n^2(X_1,\cdots,X_n) < \infty$ for every sequence $X_1,\cdots,X_n,\cdots$. The recursion considered is

$$X_{n+1} = X_n - a_n(\cdot)c_n^{-1}(Y(X_n - c_n) - Y(X_n + c_n))$$

where $Y(X_n \pm c_n)$ is a random variable whose conditional distribution given $X_1 = x_1,\cdots,X_n = x_n$ is the same as the distribution of $Y(X_n \pm c_n)$. It is usually assumed that $Y(X_n - c_n)$ and $Y(X_n + c_n)$ are conditionally independent, i.e., for all Borel sets A and B,

$$P(Y(X_n + c_n)\varepsilon A,\ Y(X_n - c_n)\varepsilon B \mid X_n) = P(Y(X_n + c_n)\varepsilon A \mid X_n)\ P(Y(X_n - c_n)\varepsilon B \mid X_n).$$

Following Ref. 8 this assumption is not necessary. Let $Z(x) = Y(x) - M(x)$ and denote $M(X_n - c_n) - M(X_n + c_n)$ by M_n and $Z(X_n - c_n) - Z(X_n + c_n)$ by Z_n then the recursion becomes

$$X_{n+1} = X_n - a_n(\cdot)\, c_n^{-1}(M_n + Z_n)$$

As a result of the assumptions on the distribution of $Y(X_n \pm c_n)$ and the fact that $EY(x) = M(x)$, $EZ(x) = 0$ for all x and the conditional distribution Z_n given $X_1 = x_1, \cdots, X_n = x_n$ is the same as the distribution of $Z(X_n - c_n) - Z(X_n + c_n)$. This implies that $E(Z_n | Z_1, \cdots, Z_{n-1}) = E(Z_n | X_1, \cdots, X_n) = 0$ wpl.

The first four assumptions deal with the behavior of the regression function $M(x)$ and are similar to those made by Kiefer and Wolfowitz.[5] Intuitively, these conditions require that outside a small neighborhood of the location of the maximum, say θ, the slope is bounded away from zero, $M(x)$ should have no plateaus and be almost quadratic when it is expanded about θ. Moreover, in the entire parameter space $M(x)$ must be "smooth enough" so that radical slope changes cannot cause the recursive procedure to oscillate indefinitely. More explicitly:

(1) $M(x)$ is a Borel measurable function which has a unique maximum at $x = \theta$ and for $0 < t_0 < t_1 < t_2 < \infty$,

$$t_1 < \inf_{0 < \varepsilon \le t_0} |x - \theta| < t_2 \quad (x - \theta)(M(x - \varepsilon) - M(x + \varepsilon))/\varepsilon > 0 \ .$$

In addition, for all x and suitable D_1 and D_2, $|M(x+1) - M(x)| \le D_1 + D_2|x|$.

(2) For all x, $M(x) = a_0 - a(x - \theta)^2 + \delta(x, \theta)$ where a_0 is some real number, $a > 0$, and $\delta(x, \theta) = \delta(|x - \theta|^2)$ as $x - \theta \to 0$.

(3) For some $c_0 > 0$, there exists positive constants K_1 and K_2 s.t. for any x and for all c for which $0 < c \le c_0$,

$$K_1(x - \theta)^2 \le (x - \theta)\,(M(x - c) - M(x + c))\, c^{-1} \le K_2(x - \theta)^2$$

(4) For every $\varepsilon > 0$, there exists $c_\varepsilon > 0$ s.t. for all c satisfying $0 < c < c_\varepsilon$ and all x satisfying $|x - \theta| < c$, $|\delta(x - c, \theta) - \delta(x + c, \theta)|\, c^{-1} \le \varepsilon |x - \theta|$.

The above assumptions limit the class of regression functions for which one has convergence in law. In addition, assumptions on the additive noise $Z(x)$ are needed. The first requires the variance of $Z(x)$ to be bounded for any x and the second condition which is virtually a uniform integrability condition upon the sequence $\{Z^2(X_n)\}$, establishes the Lindeberg condition.

(5) $\sup_x EZ^2(x) = s < \infty$ and $\lim_{\substack{x\to\theta \\ a\to 0}} E(Z(x-a) - Z(x+a))^2 = \sigma^2$.

In case $Z(x-a)$ and $Z(x+a)$ are uncorrelated we can replace the latter equation with $\lim_{x\to\theta} EZ^2(x) = \sigma^2/2$.

(6) $\sup_{|x-\theta|<\varepsilon} E|Z(x)|^{2+v} < \infty$ for some $\varepsilon > 0$ and $v > 0$.

The last condition places constraints upon $A(X_1, \cdots, X_n)$ in order that it be mean-square consistent and bounded.

(7) (a) $0 < A' = \inf_{\{X_i, 1\le i\le n\}} A_n(\cdot) \le \sup_{\{X_i, 1\le i\le n\}} A_n(\cdot) = A'' < \infty$ wpl

(b) $\lim_n \inf_{\{X_i, 1\le i\le n\}} A_n(\cdot) \ge A_1 > 0$ wpl

(c) $E(A_n(\cdot) - A)^2 \to 0$ as $n \to \infty$ s.t. $A_1 \le A \le A''$.

Theorem 1: Suppose that conditions (1) to (7) are satisfied. Let $AK_1 > A_1K_1 > 1/2$ and let $a_n(X_1, \cdots, X_n) = A(X_1, \cdots, X_n)/n$. Let $\{c_n\}$, $\{a_n(\cdot)\}$ be sequences of positive numbers satisfying $\Sigma A_n(\cdot)/n = \infty$ and $\Sigma a_n^2(\cdot) c_n^{-2} < \infty$ for all sequences $X_1, \cdots, X_n, \cdots$ and $c_n \to 0$ and $n \to \infty$, where $c_m/c_{m+1} = 1 + \varepsilon_m m^{-1}$ where $\varepsilon_m \to 0$ as $m \to \infty$. Then $\sqrt{n}\, c_n(X_n - \theta)$ is asymptotically normal with mean 0 and variance $\sigma^2A^2/8\,\alpha A - 1)$.

Proof: See Appendix of Reference 1.

B. Adaptive Threshold Detection Using Stochastic Approximation

Because the stochastic approximation algorithms are so computationally attractive, their use in conjunction with adaptive systems such as variable threshold detectors for block data is a natural outgrowth. In this example, we illustrate the application of both the RM and the KW procedures to a variable threshold detector investigated in Ref. 4 and based upon a modified version of Mosteller's k-sample slippage test.[6]

The detector is a two-sample binary detector with a reference sample $\underline{Y}_n = (y_1, \cdots, y_n)$ of size n and test sample $\underline{X}_n = (x_1, \cdots, x_n)$. The reference sample is taken under the null hypothesis of no signal present and the nth order statistic is extracted and denoted by Y_{max}. The slippage statistic R is defined to be the number of X_i which are greater than or equal to Y_{max}. More succinctly

$$R = \sum_{i=1}^{n} \{X_i \ge Y_{max}\}$$

and the test of hypothesis is

$$H_0 : \prod_{i=1}^{n} F(x_i) \quad \text{i.e., no signal present}$$

$$H_1 : \prod_{i=1}^{n} G(x_i) \quad \text{e.g., DC signal present.}$$

The hypothesis testing problem may also represent two grey levels in a picture processing problem. We assume G(x) is stochastically larger than F(x), so H_0 is rejected if $R \geq R_0$ where $R_0 \in \{0, 1, \cdots, n\}$. The discrete distribution of R under H_0 due to Mosteller is:

$$1 - F_0(r) = \binom{n}{r} / \binom{2n}{r}, \quad r = 0, 1, \cdots, n$$

and under H_1, this distribution becomes

$$1 - F_1(r) = \binom{n}{r} \int (1 - G(x))^r G^{n-r}(x)\, dF^n(x)$$

To simulate noise across picture edges, F(x) is assumed to be uniform $[-1/2, 1/2]$ and G(x) uniform $[-1/2+a, 1/2+a]$ where $\underline{a}$ is the slippage parameter representing the difference in the grey levels.

The threshold R_0 which minimizes the total probability of error P_e, depends upon $\underline{a}$ and the a priori probabilities of H_0 and H_1 which are assumed to be equally likely. P_e vs. R_0 is plotted in Fig. 1 for several values of slippage. Since R_0 is an integer these curves are obtained by randomizing the decision rule where the threshold T is interpreted as

$$T = \begin{cases} [T] \text{ wp } T - [T] \\ [T] \text{ wp } 1 - (T - [T]) \end{cases}$$

where $[x \cdot y] = x$ and $[x \cdot y] = x + 1$. To operate in the vicinity of the optimum threshold R_0, a supervised stream of binary digits is transmitted and used to adjust the threshold close to its optimum level via stochastic approximation. The random process observed is $Y(x) = I_{\{\text{error}, x\}}$ where the event {error, x} = occurrence of an error when x is the threshold. Note that because of the peaked nature of the curves in Fig. 2 that the theorem in Section A may not be applied although the KW procedure[4] does apply and guarantees mean-square convergence. Burkholder[2] establishes the asymptotic normality of the KW procedure for the class of regression functions $\mathcal{M}_\theta = \{M(x) : M : R \to R \text{ s.t. } M(x) \text{ increases (decreases) for } x < \theta \text{ and decreases (increases) for } x > \theta\}$ which includes

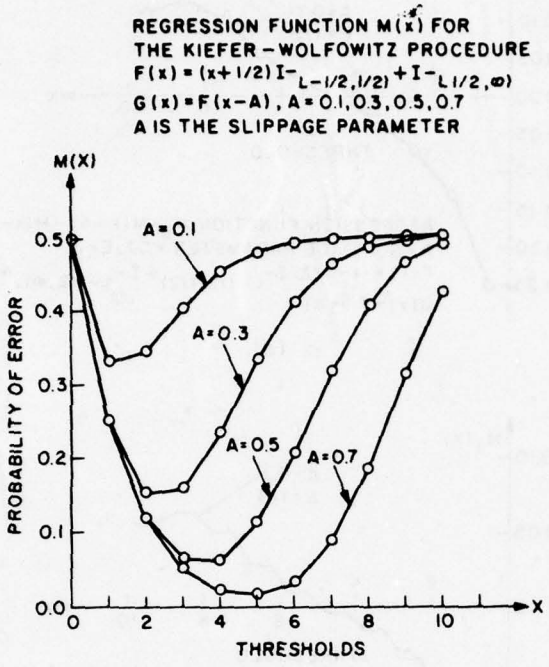


Fig. 1. Plot of the K-W regression function for the adaptive threshold detector.

the regression function under study. For $M(x) \in \mathcal{M}_\theta$ he establishes the existence of a function $\mu(\varepsilon): R$ s.t. for any $\varepsilon > 0$, $(x - \mu(\varepsilon))(M(x-\varepsilon) - M(x+\varepsilon)) > 0$, $x \neq \mu(\varepsilon)$ and $|\mu(\varepsilon) - \theta| < \varepsilon$ for each $\varepsilon > 0$. Since $P_e = M(x) = \frac{1}{2} + \sum_{i=0}^{n} (\gamma_i - \gamma_{i-1})(x - i) I_{\{x \geq i\}}$ where γ_i is the slope of $M(x)$ for $x \in (i, i+1)$, one can show with further analysis that $\mu(\varepsilon) = R_0 + \varepsilon(\gamma_u + \gamma_l)/(\gamma_l - \gamma_u)$ where γ_u and γ_l are the slope above and below the break point R_0. Using the standard KW recursion with $c_n = \varepsilon$, $a_n = A/n$, $2a_1 = M'(\mu(\varepsilon) - \varepsilon) - M'(\mu(\varepsilon) + \varepsilon)$, $2\sigma^2 = \mathrm{Var}(Y(\mu(\varepsilon) - \varepsilon)) + \mathrm{Var}(Y(\mu(\varepsilon) + \varepsilon))$, Burkholder shows the $\sqrt{n}\,(X_n - \mu(\varepsilon)) \sim N(0, 2\sigma^2 A^2/(4a_1 A - 1))$. Note that X_n does not converge to R_0, but to $\mu(\varepsilon)$ located within a distance ε of R_0. Simulation of this procedure using a restricted parameter space, i.e., the recursion is restricted to a few integers above and below R_0, shows that three thousand samples or less is sufficient to locate R_0.

The KW procedure requires two samples per iteration where the RM procedure requires only one. By choosing $Y_n = I_{\{\text{error}, x+\varepsilon\}} - I_{\{\text{error}, x-\varepsilon\}}$, $\theta = \mu(\varepsilon)$, $\sigma^2 = \mathrm{Var}\, Y_n(\mu(\varepsilon))$, and applying the theorem of Section 3 of Ref. 1 with constant gain coefficient A one concludes that $\sqrt{n}\,(X_n - u(\varepsilon)) \sim N(0, \sigma^2 A^2/(2a_1 A - 1))$ where $a_1 = \gamma_u + \gamma_l$. Thus the RM procedure may be applied as well and simulation using a restricted

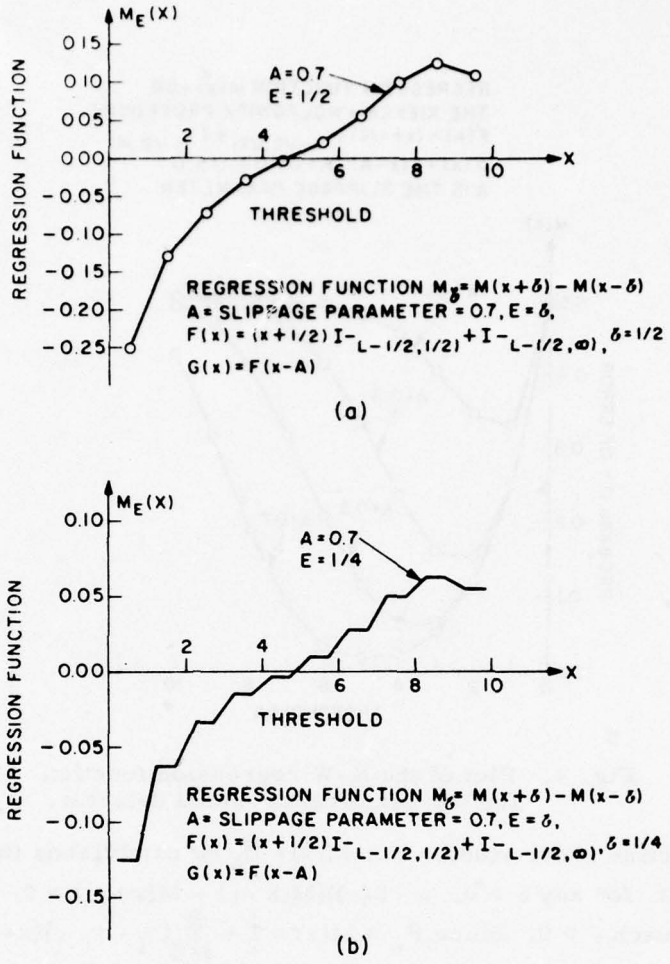


Fig. 2. R-M regression function for the variable threshold detector.

parameter space indicate it performs better than the KW procedure. The regression function for $\varepsilon = 1/2$ and $\varepsilon = 1/4$ for $a = 0.7$ is plotted in Figure 2.

The rate of convergence of both of the above procedures will deterioriate in the presence of Burst Noise. Consider what happens if F(x), instead of being uniform, is a mixture distribution of the form

$$F(x) = (1-\delta)\left[(x+1/2) I_{[-1/2,\ 1/2]}(x) + I_{[1/2,\ \infty]}(x)\right] +$$

$$\delta\left[\frac{(x+B)}{2B} I_{[-B,\ B]}(x) + I_{[B,\ \infty]}(x)\right]$$

where B is the order of 10^3 and δ is the order of 10^{-1}. Since the reference sample is

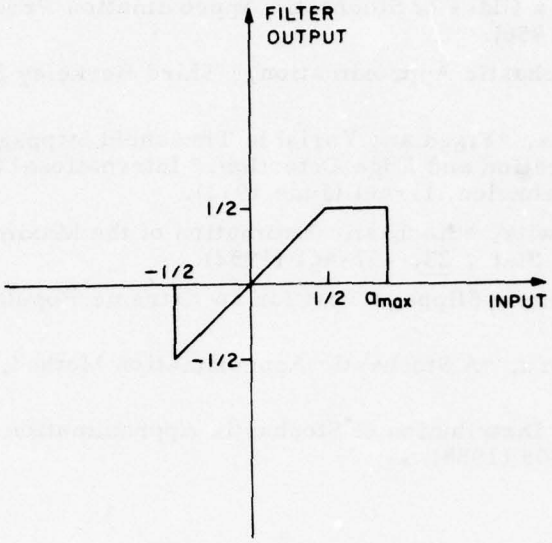


Fig. 3. Filter used in variable threshold detector.

used over many decision intervals, if Y_{max} is due to a Burst Noise sample, all decisions based on this reference sample would be determined solely by the Burst Noise. Therefore, great pains must be taken to eliminate any Burst Noise samples from the reference sample and this can easily be done with proper filtering during the period when this sample is being collected. During the decision interval a filter similar to that in Fig. 3, where a_{max} is the maximum anticipated slippage parameter will improve convergence. One must be careful to note that introduction of a filter will change the shape of the regression curves and sometimes the optimum threshold. Optimum filter design is a problem that must be tailored to a specific application and does not concern us here. In simulation an interesting phenomenon was observed. Since the regression function for the RM procedure for $\varepsilon = 1/4$ looks like a sequence of light limiters stacked upon one another (Fig. 2), one suspects that the RM procedure may get temporarily "hung up" on the wrong "ledge." This has been observed for the small sample size mentioned. However, the Burst Noise when introduced acts as either, exercising the system and pushing the recursion to the correct "ledge." In this case one could view the regression function as having its own "built in" filter.

Joint Services Technical Advisory Committee
F44620-74-C-0056

P. Kersten and L. Kurz

REFERENCES

1. J. Evans, P. Kersten and L. Kurz, "Robustized Recursive Estimation with Applications," Information Sciences, to appear.

2. D. L. Burkholder, "On a Class of Stochastic Approximation Processes," Ann. Math. Stat., 27, 1049-1059 (1956).

3. A. Dvoretzky, "On Stochastic Approximation," Third Berkeley Symposium, 1, 39-55 (1956).

4. R. Jakubek and L. Kurz, "Fixed and Variable Threshold Slippage Procedures for Block Pattern Classification and Edge Detection," International Symposium on Information Theory, Ashkelon, Israel (June 1973).

5. J. Kiefer and J. Wolfowitz, "Stochastic Estimation of the Maximum of a Regression Function," Ann. Math. Stat., 23, 457-461 (1952).

6. F. Mosteller, "A k-Sample Slippage Test for an Extreme Population," Ann. Math. Stat., 19, 58-65 (1948).

7. H. Robbins and S. Monro, "A Stochastic Approximation Method," Ann. Math. Stat., 22, 400-407 (1951).

8. J. Sacks, "Asymptotic Distribution of Stochastic Approximation Procedures," Ann. Math. Stat., 29, 373-405 (1958).

RECURSIVE FACTOR ANALYSIS METHODS IN FEATURE EXTRACTION PROBLEMS

L. Kurz and C.S. Yoon

The problem of data reduction, feature extraction and pattern recognition using factor analysis techniques were studied in References 1 and 2. In particular, stochastic approximation techniques were applied to an iterative version of the principal factor method after estimating the unique factors by ad hoc rules. Subsequently, this procedure was applied to the classification of line drawings by applying it to a precoded data.

The basic problem considered here is to find the estimates of the factor loading matrix, $\underline{A}$, and unique matrix, $\underline{D}^2$, based on all of n partitioned observations with the assumption that the underlying statistical parameters undergo slow change. This assumption is realistic for many useful data sets such as sleep-stage electro-encephalograms or record of long duration survey of earth resources. In particular, maximum likelihood recursive methodology is developed to find intuitively appealing estimates of unique factors.

One of the strategies to handle the problem is to divide the data evenly into k sections and to process each section one at a time. It is assumed that the sample covariance $Q(T_n)$, based on each section of the data T_n, is given by

$$Q(T_n) = Q(T_1, T_2, \cdots, T_k) + \alpha_n$$

$$= Q + \alpha_n \qquad n \leq k$$

where α_n is a p x p random matrix, and the adaptive sample covariance matrix $Q(T_1, \cdots, T_n)$, based on the data $(T_1, T_2, \cdots, T_n)$ is given by

$$Q_n = Q(T_1, \cdots, T_n) = Q + \beta_n$$

where β_n is a pxp random matrix. It is intuitively appealing to conclude that

$$E[\alpha_n]_{ij} > E[\beta_n]_{ij}$$

where $[\alpha_n]_{ij}$ and $[\beta_n]_{ij}$ are the ij-th elements of α_n and β_n, respectively. The approach to this partitioned data is to estimate $\underline{D}^2$ by the maximum likelihood method and to estimate $\underline{A}$ by a stochastic approximation procedure.

A. Estimation of the Factor Loading Matrix by Stochastic Approximation

The problem considered in this section is to find the estimates $\hat{\underline{A}}$ and $\hat{\underline{D}}^2$ based on all of n observations by iterative procedures. There might be several ways to estimate them. Two new methods are proposed in this section. The first method utilizes iterative algorithm to obtain a sample covariance matrix based on the whole data and applies

Jennrich's algorithm[3] to find the estimate, $\underline{\hat{A}}$, of $\underline{A}$. In the second method, Robbins-Monro type stochastic approximation algorithm is used to estimate $\underline{D}^2$ and $\underline{A}$ after the initial value $\underline{D}_1^2$ and $\underline{A}_1$ are chosen. The algorithm based on the stochastic approximation has been proved to converge and this convergence can be accelerated using various schemes.

1. Method 1

The sample mean vector M_1 based on the first section of data is

$$M_1 = M(T_1) = \frac{1}{n_1} \sum_{i=1}^{n_1} x_{ij} \tag{1}$$

where n_1 is the number of observation in the first section T_1.

$i = 1, 2, \cdots, n$ represents measurement and

$j = 1, 2, \cdots, N_1$ represents individual samples to be classified.

The mean vector of Eq. (1) is updated every time when a new section of data arrives at the processor. At the second stage the mean vector is given by

$$M_2 = M(T_1, T_2) = \frac{1}{n_1 + n_2} \{n_1 M(T_1) + n_2 M(T_2)\} \tag{2}$$

and the updated mean vector is an unbiased estimate of the population mean. Thus, the updated mean M_ℓ is calculated using the previous mean $M_{\ell-1}$ by

$$M_\ell = M(T_1, T_2, \cdots, T_{\ell-1}, T_\ell) = \frac{1}{n_1 + n_2 + \cdots, n_\ell} \left\{(n_1 + n_2 + \cdots, n_{\ell-1} + n_\ell M(T_\ell)\right\} \tag{3}$$

$\ell = 2, 3, \cdots, k$

The sample covariance matrix, $Q(T_1)$, based on the section T_1 is

$$Q_1 = Q(T_1) = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (\bar{x}_i - M_1)(\bar{x}_i - M_1)' = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} \bar{x}_i \bar{x}_i' - \frac{n_1}{n_1 - 1} M_1 M_1' \tag{4}$$

In general, the sample covariance matrix based on K observations is given by

$$Q = \frac{1}{k - 1} \sum_{i=1}^{K} x_i x_i' - \frac{k}{k - 1} M M' \tag{5}$$

and it can be calculated iteratively by using

$$Q_{n+1} = Q(T_1, T_2, \cdots, T_n, T_{n+1})$$

$$= \frac{1}{N_{n+1} - 1} \sum_{i=1}^{N_{n+1}} x_i x_i' - \frac{N_{n+1}}{N_{n+1} - 1} M_{n+1} M_{n+1}'$$

$$= \frac{N_{n+1}}{N_{n+1} - 1} \left[\frac{N_n - 1}{N_{n+1}} Q_n + \frac{1}{N_{n+1}} \sum_{i=N_n+1}^{N_{n+1}} x_i x_i' + \frac{N_n}{N_{n+1}} M_n M_n' - M_{n+1} M_{n+1}' \right] \quad (6)$$

It is easy to show the unbiasedness of the updated covariance given the algorithm in Equation (6). Namely,

$$E\{Q_1\} = \sum$$

$$E\{Q_2\} = \frac{N_2}{N_2 - 1} \left\{ \frac{N_1 - 1}{N_2} Q_1 + \frac{1}{N_2} \sum_{i=N_1+1}^{N_2} x x' + \frac{N_1}{N_2} M_1 M_1' - M_2 M_2' \right\}$$

$$= \frac{1}{N_2 - 1} E\left\{ \sum_{i=1}^{N_2} x_i x_i' \right\} - \frac{N_2}{N_2 - 1} E\left\{ M_2 M_2' \right\}$$

$$= \frac{1}{N_2 - 1} \left\{ N_2 \sum + N_2 \mu \mu' \right\} - \frac{N_2}{N_2 - 1} \left\{ \frac{\sum}{N_2} + \mu \mu' \right\}$$

$$= \sum$$

By using the iterative algorithm in Eqs. (6) and (3), a significant computational efficiency can be achieved, especially if storage capability of the processor is restricted.

With the unique factors decided by one of the ad hoc methods discussed in Harman[4] and the final updated covariance matrix, Q_k, the factor loading matrix, $\underline{A}$ based on the whole observations can be calculated by one of the estimation methods such as the principal factor method, centroid method, minres method or maximum likelihood method.

2. Method 2

When the first section T_1 arrives, the sample covariance Q_1 is calculated and the unique factor $\underline{D}_1$ is estimated by the maximum likelihood estimation method. While $\underline{D}_1$ is estimated, $\underline{A}_1$ is calculated in such a way that

$$(Q_1 - \underline{D}_1^2) = \underline{A}_1 \underline{A}_1'$$

Everytime when the new section of observations arrives, the sample covariance is updated and the matrix $\underline{A}$ is also updated until the final factor loading matrix, based on

all observations, is obtained. The stochastic approximation procedure is applied to obtain the final matrix $\underline{A}$ with $\underline{A}_1$ and $\underline{D}_1$ as initial values. One may develop three sets of recursive procedures to estimate $\underline{A}$ and $\underline{D}$ from the relation,

$$(\underline{Q} - \underline{D}^2)\underline{A} - \underline{A}\underline{J} = 0 \tag{6a}$$

where

$$\underline{J} = \underline{A}'\underline{A}$$

which was derived using the principal factor method, and another relation

$$\underline{D}^2 = \text{diag}\,(\underline{Q} - \underline{A}\underline{A}') \tag{7}$$

which was obtained from the maximum likelihood estimation method. The usual RM type recursions

$$\underline{A}_{n+1} = \underline{A}_n + \frac{a}{n}\left[(\underline{Q}_n - \underline{D}_n^2)\underline{A}_n - \underline{A}_n\underline{J}_n\right] \tag{8a}$$

$$\underline{D}_{n+1}^2 = \underline{D}_n^2 + \frac{b}{n}\left[\text{diag}\,(\underline{Q}_n - \underline{A}_n\underline{A}_n' - \underline{D}_n^2)\right] \tag{8b}$$

are not suitable in this problem because

$$(\underline{Q}_1 - \underline{D}_1^2)\underline{A}_1 - \underline{A}_1\underline{J}_1 = 0 \tag{9}$$

Since it was assumed that the parameters of distribution of the data are slowly varying, one may consider that

$$\underline{Q}_{n+1} = \underline{Q}_n + \underline{\delta}_n + \varepsilon \tag{10}$$

$$n = 1, 2, 3, \cdots$$

where $\underline{\delta}_n$ is a zero-mean error matrix and each element

$$[\varepsilon]_{ij} << [Q_n]_{ij}$$

$\underline{Q}_n$ will be replaced by $\underline{Q}_{n+1}$ in the improvement part of the iteration. It is possible then to generate the RM type recursions to update a factor loading matrix and a unique factor matrix.

At first, the RM type recursive equation can be developed from the relations obtained by the principal factor method. Suppose that an updated covariance matrix, $\underline{Q}_{n+1}$ is available, then a factor loading matrix, $\underline{A}_n$ is updated by

$$\underline{A}_{n+1} = \underline{A}_n + \frac{C_1}{n}\left[(\underline{Q}_{n+1} - \underline{D}_n^2)\underline{A}_n - \underline{A}_n\underline{J}_n\right] \tag{11a}$$

where $\{C_1\}$ is a gain sequence, and the unique factor matrix is

$$\underline{D}^2_{n+1} = \text{diag}(\underline{Q}_{n+1} - \underline{A}_{n+1}\underline{A}'_{n+1}) \tag{11b}$$

Secondly, a recursive version can be constructed from the equation developed from the maximum likelihood equation and

$$\underline{D}^2_{n+1} = \underline{D}^2_n + \frac{C_2}{n}[\text{diag}\,(\underline{Q}_{n+1} - \underline{A}_n\underline{A}'_n - \underline{D}^2_n)] \tag{12a}$$

$$(\underline{Q}_{n+1} - \underline{D}^2_{n+1})\underline{A}_{n+1} = \underline{A}_{n+1}\underline{J}_{n+1} \tag{12b}$$

are obtained. Since eigenvectors should be calculated repeatedly in Eq. (21b), this set of equations is not computationally useful.

In view of that fact, a new procedure is used. The recursive procedures generated from both methods can be used to update two matrices. The unique factor matrix is updated by

$$\underline{D}^2_{n+1} = \underline{D}^2_n + \frac{C_2}{n}[\text{diag}\,(\underline{Q}_{n+1} - \underline{A}_n\underline{A}'_n - \underline{D}^2_n)] \tag{13a}$$

then the factor loading matrix is improved by

$$\underline{A}_{n+1} = \underline{A}_n + \frac{C_1}{n}[(\underline{Q}_{n+1} - \underline{D}^2_{n+1})\underline{A}_n - \underline{A}_n\underline{J}_n] \tag{13b}$$

The convergence of sequence, $\underline{A}_n$ generated by the algorithm in Eq. (11) will be discussed in Theorem 1 and the one generated by algorithm Eq. (13) will be discussed in Theorem 2.

Theorem 1: Suppose that the observed random variables $\underline{X}$ have bounded covariance matrices $\underline{Q}_1, \underline{Q}_2, \cdots$, and each of them is a covariance matrix of a factor analysis model. Let us choose $\underline{A}_1$ and $\underline{D}^2_1$ such that

$$(\underline{Q}_1 - \underline{D}^2_1) = \underline{A}_1\underline{A}'_1$$

The sequence of estimators $\{\underline{A}_n\}$ generated by the algorithm Eq. (11a) converges to the $\underline{A}$ satisfying Eq. (3) with probability 1.

Proof: See Reference 5.

Theorem 2: Suppose that the assumptions in Theorem 1 hold and $\underline{D}^2_n$ generated by the algorithm in Eq. (13a) has an asymptotically bounded variance, then the sequence of estimators $\{\underline{A}_n\}$ generated by the algorithm in Eq. (13b) converges to the matrix $\underline{A}$ satisfying Eq. (3) with probability 1.

Proof: See Reference 5.

Variance of $\underline{D}_n^2$

Since the variance of $\underline{D}_n$ is an important factor to prove convergence of the algorithm in Theorem 2 and $\underline{D}_n^2$ is generated iteratively by the algorithm in Eq. (13a), the study of the asymptotic behavior of the variance of $\underline{D}_n$ is of interest. This behavior yields information about the quality of the estimator.

Since $\underline{D}_n^2$ and the term in the brackets of the algorithm in Eq. (13a) are diagonal matrices, they can be considered as vectors. The vector version of Sacks' asymptotic normality theorem[6,7] could be used to find the asymptotic variance since the form of the algorithm fits Sacks' iteration scheme. To avoid unnecessary complexities, a scalar version for each component will be used. The components of $\underline{D}_n^2$, d_{ni}, are iteratively computed by

$$d_{n+1,i} = d_{n,i} + \frac{b_{n,i}}{n}\left\{(q_{n+1})_{ii} - \sum_{j=1}^{K} (a_n)_{ij}^2 - D_{n,i}\right\} \tag{14}$$

$$i = 1,2,3,\cdots,N \qquad n = 1,2,\cdots,$$

where $(q_{n+1})_{ii}$ is the i-th diagonal term of the covariance matrix $\underline{Q}_{n+1}$. Thus, n scalar estimators are generated simultaneously, yielding the components of the vector estimator.

Theorem 3: Suppose that the i-th component of $\underline{D}_n^2$ is estimated by the algorithm in Eq. (14) with initial values $\underline{D}_1^2$ and $\underline{A}_1$ obtained by one of factor loading estimation methods. If the following conditions are satisfied let

$$Y_n(d_{n,i}, d_i) = M(d_{n,i}, d_i) + Z_n(d_{n,i}, d_i)$$

where $E[Z(d_{n,i}, d_i)] = 0$ and $M(d_{n,i}, d_i)$ has a unique root at $d_{n,i} = d_i$.

Condition (1)

$(d_{n,i} - d_i)\, M(d_{n,i} - d_i) > 0$ for all $d_{n,i} \neq d_i$

Condition (2)

$K|d_{n,i} - d_i| \leq |M(d_{n,i}, d_i)| \leq K_1|d_{n,i} - d_i|$

Condition (3)

$M(d_{n,i}, d_i) = \alpha_1(d_{n,i} - d_i) + \delta(d_{n,i}, d_i)$

Condition (4)

a) $\sup_{d_i'} E\{Y(d_i') - M(d_i')\}^2 < \infty$

b) $\lim_{d_i' \to d_i} E\{Y(d_i') - M(d_i')\}^2 = \sigma^2$

Condition (5)

$$\sup_{|d_i' - d_i| < \varepsilon} E\{Y(d_i') - M(d_i')\}^{2+\nu} < \infty$$

for some $\varepsilon > 0$, $\nu > 0$

then the sequence of estimators, d_{ni}, converges to d_i with probability one and its asymptotic variance is given by

$$\lim_{n \to \infty} \text{var}(d_{ni}) = \frac{\sigma^2 b_i^2}{n(2b_i - 1)}$$

Proof: See Reference 5.

Thus, two sets of iterative algorithms have been developed to extract a factor loading matrix from a continuous flow of data. The well developed stochastic approximation algorithm was applied to maximum-likelihood estimation of the factor loading matrix. By proper selection of gain constants and weighing functions, efficient operation of the algorithms is assured.

Joint Services Technical Advisory Committee
F44620-74-C-0056

L. Kurz and C.S. Yoon

REFERENCES

1. J. Aron and L. Kurz, "Feature Extraction by Recursive Factor Analysis," Proc. 1973 Princeton Conference on Information Sciences and Systems (March 1973).
2. L. Kurz and C.S. Yoon, "Classification of Line Drawings by Factor Analysis of Precoded Data," Progress Report No. 40 to JSTAC, Polytech. Inst. of New York, Report No. R-452.40-75, 609-611 (November 1975).
3. R.I. Jennrich and S.M. Robinson, "A Newton-Raphson Algorithm for Maximum Likelihood Factor Analysis," Psychometrike, 34, 111-123 (1969).
4. H.H. Harman, "Modern Factor Analysis," (Chicago: Univ. of Chicago Press, 1967).
5. C.S. Yoon, "Encoding, Reduction and Classification of Two-Dimensional Data by Factor Analysis," Ph.D. Dissertation, Polytech. Inst. of New York (June 1976).
6. R.D. Martin and C.J. Masreliez, "Robust Estimation via Stochastic Approximation," IEEE Trans. on Information Theory, IT-21, No. 3, 263-271 (1975).
7. J. Sacks, "Asymptotic Distribution of Stochastic Approximation Procedures," Ann. Math. Stat., 35, 590-599 (1964).

NONPARAMETRIC PROBABILITY DENSITY ESTIMATION

I.M. Habib and L. Kurz

The estimation of probability density functions (p.d.f.'s) and related problems had always been of major interest in processing of computed data. The problem has additional attraction because once it is solved, related problems of channel condition monitoring, design of adaptive detectors, selection of optimum scores for nonparametric detectors become easily tractable. The determination of an optimal in some sense p.d.f. that fits a set of independent identically distributed samples (i.i.d.) has been considered by many researchers.[1-16] The approach taken by the investigators was to obtain robust and nonparametric procedures for estimation of unknown p.d.f. However, these methods resulted in estimators which fit a certain p.d.f. quite well and do not fit other p.d.f.'s. Essentially, these procedures are nonparametric but not robust.

The method described here is based on the philosophy suggested by Cencov.[9] There are two basic differences in the approach taken. Instead of unspecified orthonormal sets, three specific orthonormal sets of functions are used which result in excellent estimators for three classes of distributions: near-gaussian, clipped impulsive noise and skewed. The choice of the estimation procedure in a particular application results in efficient use of computation time and minimum number of terms in the expansion for a given mean-square error of estimation. The procedures presented compare favorably with the most efficient published results.

A. Formulation of the Estimation Problem

The estimate of the unknown p.d.f., f(x), based on N i.i.d. samples is of the form

$$\hat{f}_N(x) = \sum_{i=1}^{r} a_{iN}\theta_i(x) \tag{1}$$

where

a_{iN} , $i = 1, \cdots, r$ are the expansion coefficients

$\theta_i(x)$, $i = 1, \cdots, r$ are the appropriately preselected sets of orthonormal functions

If the error

$$e = \int_x [f(x) - \sum_{i=1}^{r} a_{iN}\,\theta_i(x)]^2\,dx \tag{2}$$

is to be minimized then the coefficients a_{iN} must satisfy

$$E[\theta_i(x) - a_{iN}] = 0 \quad , \quad i = 1, \cdots, r \tag{3}$$

To estimate a_{iN}, a Robbins-Monro one-dimensional stochastic approximation algorithm is used. In particular, the recursion relation is

$$a_{iN}^{(k+1)} = a_{iN}^{(k)} + \gamma_k(\theta_i(x^{(k)}) - a_{iN}^{(k)}) \quad , \quad i = 1, \cdots, r \tag{4}$$

where

$$\gamma_k = \frac{1}{k+1}$$

and $a_{iN}^{(k)}$ converges to a_{iN} in the mean-square sense and with probability one.

B. Gaussian and Near-Gaussian Distributions

For this class of distributions, the best set of orthogonal functions is the set of Hermite polynomials defined by

$$H_i(x) = (-1)^i e^{x^2} \frac{d^i}{dx^i}(e^{-x^2}) \tag{5}$$

with the orthogonality equations

$$\int_{-\infty}^{\infty} e^{-x^2} H_i(x)\, H_j(x)\, dx = 2^i i!\, \pi\, \delta_{ij} \tag{6}$$

where δ_{ij} is the Kronecker delta. Introducing

$$c_i = a_{iN}\, 2^i i!\, \pi \tag{7}$$

Equation (4) reduces to

$$c_i^{(k+1)} = c_i^{(k)} + \gamma_k[H_i(x^{(k)})\, e^{-(x^{(k)})^2} - c_i^{(k)}] \tag{8}$$

and the estimator of the unknown p.d.f. is

$$\hat{f}(x) = \sum_{i=1}^{r} a_{iN}\, H_i(x) \tag{9}$$

The procedure was applied to a contaminated sample of size 100 with a nominal distribution N(0, 1). The solution was carried out on a Univac digital computer and the results are shown in Figure 1. The fit is perfect and much better than using the method suggested by Good and Gaskins[14] if the comparison is based on fixed computation time.

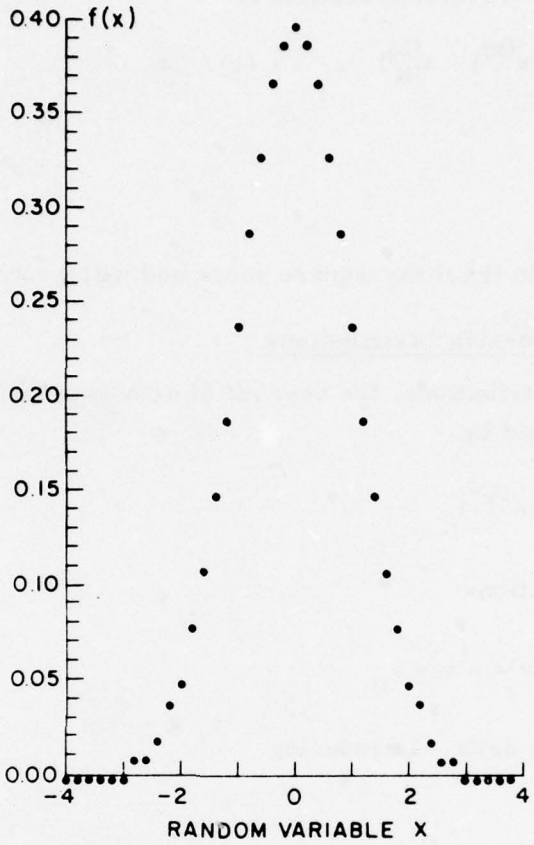


Fig. 1. p.d.f. estimator for a N(0, 1) distribution.

C. Clipped Impulsive Noise Distributions

For this class of distribution the suitable set of orthonormal function was found to be the set of prolate spheroidal wave functions[17] $\{\psi_i(c,x)\}$ which satisfy the relations

$$\int_{-\infty}^{\infty} \psi_i(c,x)\, \psi_j(c,x)\, dx = \delta_{ij} \tag{10}$$

and

$$\int_{-T/2}^{T/2} \psi_i(c,x)\, \psi_j(c,x)\, dx = \lambda_i\, \delta_{ij} \tag{11}$$

where λ_i are the eigenvalues corresponding to the eigenfunctions $\{\psi_i(c,x)\}$. Introducing

$$c_i = a_{iN}\, \lambda_i(c) \tag{12}$$

and proceeding as in Section B, the estimator of p.d.f. is then

$$\hat{f}(x) = \sum_{i=1}^{r} a_{iN} \psi_i(c, x) \tag{13}$$

The procedure was applied to a sample of size 100 drawn from the output of a hard limiter represented by

$$x_T = \int_{-T/2}^{T/2} \operatorname{sgn}[y(t)]\, dt \tag{14}$$

The input to the limiter, y(t), consisted of large variance impulsive noise represented in time by

$$p(n, T) = \frac{(\alpha T)^n \exp(-\alpha T)}{n!} \tag{15}$$

where α is the expected number of zero crossings per unit time and n is the number of zero crossings in an interval of length T. McFadden[18] gives the exact expression for $f(x_T)$ as

$$f(x_T) = \frac{\alpha e^{-\alpha T}}{2}\left[I_0(\alpha T)\sqrt{1 - \left(\frac{x_T}{T}\right)^2} + \frac{I_1\left(\alpha T\sqrt{1 - (x_T/T)^2}\right)}{\sqrt{1 - (x_T/T)^2}}\right]$$

$$+ \frac{e^{-\alpha T}}{2}\left[\delta(x_T - T) + \delta(x_T + T)\right] \quad \text{for} \quad |x_T| \leq T \tag{16}$$

$$= 0 \qquad \text{for} \quad |x_T| > T$$

where $I_0(\cdot)$ and $I_1(\cdot)$ are modified Bessel functions of the first kind, or order zero and one, respectively. For T = 10, $\alpha = .5$, c = 8, r = 8, the comparison of estimated and exact p.d.f.'s is shown in Figure 2. The discrete portion of the distribution is omitted.

D. Skewed Distributions

For this class of distributions the set of orthonormal Laguerre functions is applicable. The Laguerre functions are given by

$$L_n(\tau) = \sqrt{2p}\left[\frac{(2p)^n}{n!}\tau^n - \frac{n(2p)^{n-1}}{(n-1)!}\tau^{n-1} + \frac{n(n-1)(2p)^{n-2}}{2!(n-2)!}\tau^{n-2}\right.$$

$$- \frac{n(n-1)(n-2)(2p)^{n-3}}{3!(n-3)!}\tau^{n-3} \tag{17}$$

$$\left. + \cdots + (-1)^n\right] e^{-p\tau}$$

$$n = 0, 1, 2, \cdots \quad \text{and} \quad 0 \leq \tau < \infty$$

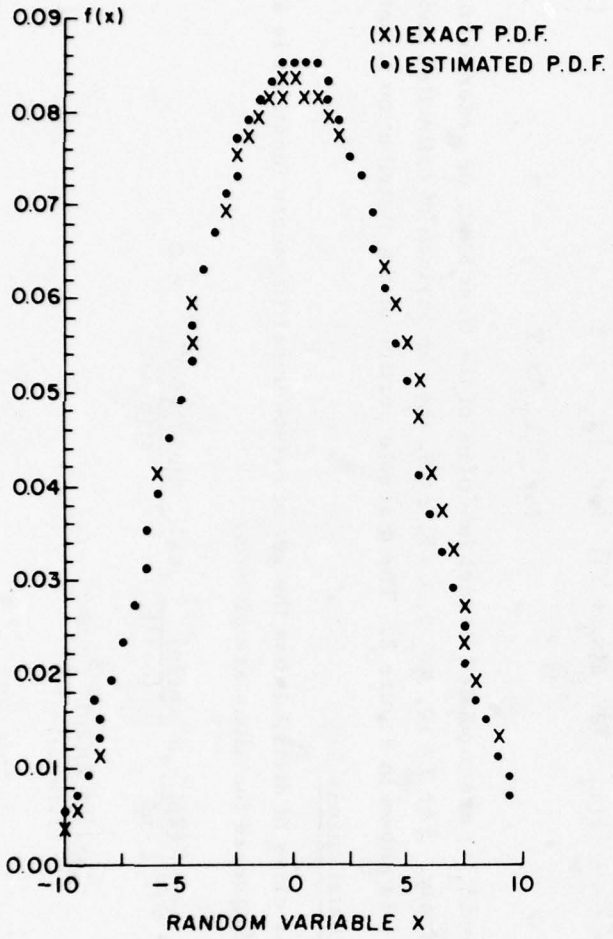


Fig. 2. p.d.f. estimator for clipped-impulsive noise.

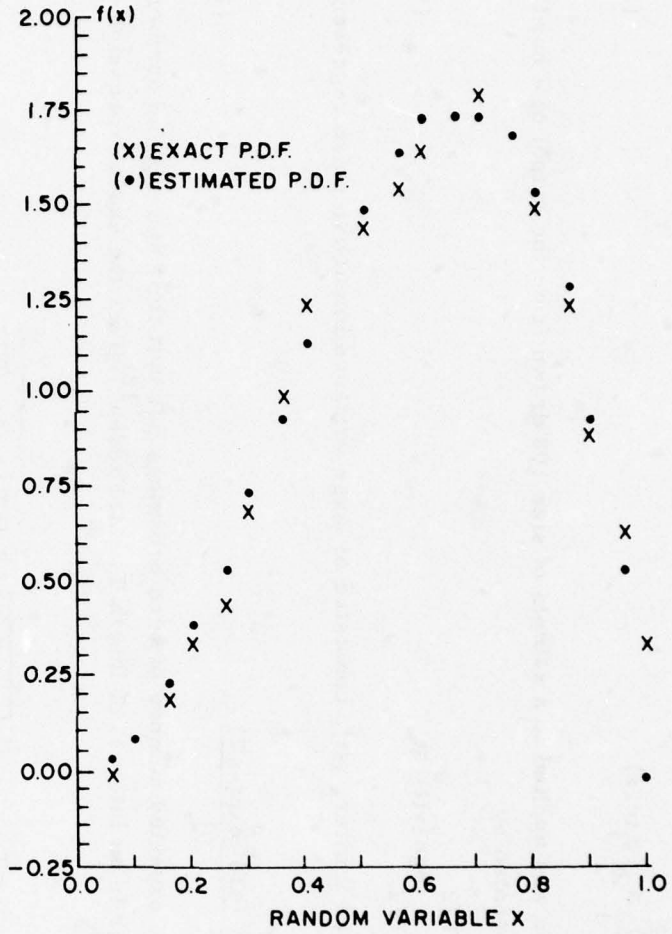


Fig. 3. p.d.f. estimator for a beta distribution.

In this case Eqs. (4) and (1) are applicable with $\theta_{iN}(x) = L_i(x)$ as an example, the estimation problem was solved for a sample size 10θ from a Beta distribution with the p.d.f. specified by

$$f(x) = \frac{(\alpha+\beta+1)!}{\alpha!\,\beta!} x^{\alpha}(1-x)^{\beta}, \quad 0 \leq x \leq 1, \ \alpha > -1, \ \beta > -1$$

$$= 0 \qquad \text{elsewhere}$$

For $\beta = 1$, $\alpha = 2$, $r = 30$ and $p = .5$, the results of the estimation are shown in Figure 3.

Joint Services Technical Advisory Committee
F44620-74-C-0056

I.M. Habib and L. Kurz

RCA GLOBCOM

REFERENCES

1. M. Rosenblatt, "Remarks on Some Nonparametric Estimates of a Density Function," Ann. Math. Stat., 27, 832-837 (1956).
2. E. Parzen, "On Estimation of a Probability Density and Mode," Ann. Math. Stat., 33, 1065-1076 (1962).
3. P. Whittle, "On Smoothing of Probability Density Functions," J. Roy. Statist. Soc. Ser. B., 20, 334-343 (1958).
4. M.S. Bartless, "Statistical Estimation of Density Functions," Sankhyā Ser. A., 25, 245-254 (1963).
5. E.F. Schuster, "Estimation of a Probability Density and Its Derivatives," Ann. Math. Stat., 40, 1187-1195 (1969).
6. J. Van Ryzin, "On Strong Consistency of Density Estimation," Ann. Math. Stat., 40, 1765-1772 (1969).
7. W.K. Murthy, "Estimation of the Probability Density," Ann. Math. Stat., 36, 1027-1031 (1965).
8. D.F. Specht, "Series Estimation of a Probability Density Function," Technometrics, 13, 409-424 (1971).
9. N.N. Cencov, "Evaluation of an Unknown Distribution From Observations," Soviet. Math., 3, 1559-1562 (1962).
10. S.C. Schwartz, "Estimation of Probability Densities by an Orthogonal Series," Ann. Math. Stat., 38, 1261-1265 (1967).
11. M.E. Tarter and R. Kronmal, "A Description of New Computer Methods for Estimating the Population Density," Proc. ACM, 22, 511-519 (1967).
12. G.S. Watson, "Density Estimation by Orthogonal Series," Ann. Math. Stat., 40, 1496-1498 (1969).
13. S.C. Schwartz, "On the Estimation of a Gaussian Convolution Probability Density," SIAM J. Appl. Math., 17, 447-453 (1969).
14. I.J. Good and R.A. Gaskins, "Nonparametric Roughness Penalties for Probability Densities," Biometrika, 58, 255-277.
15. Ya. Z. Tsypkin, "Use of the Stochastic Approximation Method in Estimating Unknown Distributions Densities from Observations," Avtomatikai Telemekhanika, 27, No. 3.
16. L. Boneva, D.G. Kendall and I. Stefanov, "Spline Transformations. Three New Diagnostic Aids for the Statistical Data Analyst," J. Roy. Statist. Soc. Ser. B
17. D. Slepian and M.O. Pollak, "Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty, I," Bell Sys. Tech. J., 40, 45-64 (January 1961).
18. J.A. McFadden, "The Fourth Product Moment of the Infinitely Clipped Noise," IRE Trans. on Information Theory, IT-4, 159-162 (December 1958).

blank

MEETING PAPERS

10th Symposium on Nondestructive Evaluation, San Antonio, Texas, April 1975

H. L. Bertoni and Y. L. Hou, "Effects of Damping in a Solid on Acoustic Beams Reflected at the Rayleigh Critical Angle"

10th Congress of International Commission for Optics, Prague, Czechoslovakia, August 1975

F. P. Mattar and M. C. Newstein, "Transverse Effects in the Propagation of Coherent Pulses"

XVIth IUGG General Assembly, Grenoble, France, September 1975

S. H. Gross, "The Onset of Instability Against Escape of an Atmosphere"

S. H. Gross and G. V. Ramanathan, "Twin Evolution of Primitive Hydrogen Atmospheres"

1975 Annual USNC/URSI Meeting, Boulder, Colorado, October 20-23, 1975

L. Birenbaum, I. T. Kaplan, W. Metlay, S. W. Rosenthal and M. M. Zaret, "Effects of 35 and 107 GHz Microwaves on the Rabbit Eye"

B. R. Cheo, "Truncation of Sheath Spiral Antennas"

L. B. Felsen, "Beams and Modes in Optical Fibers"

Fall 1975 Meeting of the Optical Society of America, Boston, Massachusetts, October 1975

M. C. Newstein and F. P. Mattar, "Self-Focusing of Coherent Optical Pulses in Resonant Absorbers"

Electro-Optics International Laser Exposition, Anaheim, California, November 1975

T. Tamir, S. T. Peng and H. L. Bertoni, "Beam and Waveguide Couplers"

National Telecommunications Conference, New Orleans, Louisiana, December 1-3, 1975

A. Kershenbaum and R. Boorstyn, "Centralized Teleprocessing Network Design"

H. Schachter, J. Franck and D. Schilling, "Three Level Delta Modulation"

21st Conference on Magnetism and Magnetic Materials, Philadelphia, Pennsylvania, December 9-12, 1975

L. Assadourian and L. M. Silber, "Crystalline Anisotropy of Cobalt Ferrite from Observations of Ferromagnetic Resonance"

55th Transportation Res. Bd. Meeting, Washington, DC, January 1976

D. J. Mayer and L. Shaw, "The Stability of Occupancy Based Library-Type Traffic Control Systems"

The American Physical Society Meeting, New York, New York, January 1976

R. Foye and S. Barone, "Classical Theory of a Neutral Magnetic Dipole in a Static Electric Field"

P. Melman and S. Barone, "Asymptotic Solution of the Dirac Equation and the Inertial Mass of an Electron"

G. Osche and B. Barone, "On the Physical Interpretation of the Dirac-Pauli Theory of a Neutral Dipole in a Static Electric Field"

Topical Meeting on Integrated Optics, Salt Lake City, Utah, January 12-14, 1976

A. A. Oliner and S. T. Peng, "Optical Slot and Strip Waveguides Formed by Metal Platings"

S. T. Peng, T. Tamir, D. W. Fradin and P. K. Cheo, "Effects of Imperfections on the Efficiency of Optical Couplers"

Operations Res. Soc. Meeting, Philadelphia, Pennsylvania, March 1976

L. Shaw and N. Wattanapanom, "Optimal Testing Times for Failure Detection when Tests Hasten Failure"

1976 Meeting of the Division of Planetary Sciences/AAS, Austin, Texas, April 1976

S. H. Gross, "Io's Ionosphere and the Sodium Clouds"

Polytechnic Institute of New York-Microwave Research Institute, Symposium on Computer Software Engineering, New York, New York, April 20-22, 1976

S. N. Mohanty and M. Adamowicz, "Proposed Measures for Evaluation of Software"

M. L. Shooman and S. Natarajan, "Effect of Manpower Deployment and Error Generation on Software Reliability"

IEEE International Conference on Communications, Philadelphia, Pennsylvania, June 1976

F. Haisch and R. Boorstyn, "Adaptive Message Switching Networks"

International Electromagnetics Symposium, University of Illinois at Chicago Circle, June 1976

N. Marcuvitz, "Quasiparticle Approach to Nonlinear and Turbulent Wave Problems"

Microwave Theory and Techniques International Symposium, Cherry Hill, New Jersey, June 1976

S. Choudhary and L. B. Felsen, "Propagation in Inhomogeneous Slab Waveguides"

L. B. Felsen, "Evanescent Waves"

National Conference on Electromagnetic Scattering, Chicago, Illinois, June 1976

L. B. Felsen, "Asymptotic Methods in Propagation and Scattering"

1976 International Symposium on Fault-Tolerant Computing, Pittsburgh, Pennsylvania, June 21-23, 1976

A. A. Sarris, M. Mor and E. J. Smith, "On Test Generation for Asynchronous Sequential Circuits"

1976 International Symposium on Information Theory, Ronneby, Sweden, June 21-24, 1976

L. Kurz and G. Pasternack, "Non-parametric Classification Using Quadratic Quantile Statistics"

1976 Inter-Society Conference on Transportation, Los Angeles, California, July 1976

M. L. Shooman, M. Horodniceanu and E. J. Cantilli, "System Safety Applied to Transportation Systems"

Microwave Mobile Communications Symposium, September 1976

R. Pennotti and R. Boorstyn, "Channel Assignments in a Cellular Mobile Telecommunications System"

6th European Microwave Conference, Rome, Italy, September 1976

L. B. Felsen, "Ray Methods for Evanescent Fields"

A. A. Oliner and S. T. Peng, "Waveguides for Integrated Optics Formed by Metal Platings"

H. Schachter, J. S. Sundresh and F. A. Cassara, "A Novel FM Detector for Suppression of Interchannel Interference"

13th IEEE Computer Society International Conference, Washington, DC, September 1976

M. L. Shooman, "Recent Developments in Software Reliability - The State of the Art"

JOURNAL ARTICLES AND CORRESPONDENCE

S. Austin and F. T. Stone, "Fabrication of Thin Periodic Structures in Photoresist: A Model," Appl. Opt., 15, 7071-7074 (April 1976).

S. Austin and F. T. Stone, "Thin Periodic Structures in Photoresist: Fabrication and Experimental Evaluation," Appl. Opt., 15, No. 9, 2126-2130 (September 1976).

S. Barone, "Spin Dependent Corrections to Commutation Relations," Phys. Rev. D, 12, 3363-3364 (November 1975).

H. L. Bertoni and T. Tamir, "Characteristics of Wedge Transducers for Acoustic Surface Waves," IEEE Trans. Sonics and Ultrasonics, SU-22, No. 6, 415-420 (November 1975).

H. L. Bertoni and T. Tamir, "Coupling Layers for Efficient Wedge Transducers," IEEE Trans. Sonics and Ultrasonics, SU-22, No. 6, 421-430 (November 1975).

F. A. Cassara, "Experimental Optimization of the Signal-to-Noise Ratio Response of the Frequency Demodulator with Feedback in the Presence of Interchannel Interferences," to be published in International Journal of Electronics.

F. A. Cassara and B. Senitzky, "A Freshman Engineering Lab," to be published in IEEE Trans. on Education.

B. R. Cheo, "Truncation of Infinite Sheath Spiral Antennas," to be published in IEEE Trans. Antennas & Propagation.

R. S. Chu and T. Tamir, "Bragg Diffraction of Gaussian Beams by Periodically Modulated Media," J. Opt. Soc. Am., 66, 220-226 (March 1976).

K. Chung and H. Eun, "Observation of Microwave Mixing in Air Glow Plasma," to be published in IEEE Trans. on Plasma Science.

E. Eun and S. H. Gross, "Ionispheric Disturbances and Gravity Waves," to be published in Geophys. Res. Lett. (1976).

J. Evans, P. Kersten and L. Kurz, "Robustized Recursive Estimation with Applications," Information Sciences, 11, No. 2, 69-92 (August 1976).

L. B. Felsen, "Complex Rays," Philips Research Reports, Special Issue in honor of C. J. Bouwkamp, 30, 187-195 (1975).

L. B. Felsen, "Alternative Representations of Green's Functions for the Helmholtz Equation," J. Opt. Soc. Am. (December 1975).

L. B. Felsen and S. Choudhary, "A New Look at Evanescent Waves," Nouv. Rev. Optique, 6, No. 5, 297-301 (1975).

L. B. Felsen and C. Santana, "Two-Dimensional and Three-Dimensional Unstable Resonator Losses by a Waveguide Analysis," Opt. Soc. Am., CLEA Special Issue (June 1976).

S. H. Gross and G. V. Ramanathan, "The Atmosphere of Io," to be published in Icarus (1976).

S. H. Gross and H. Eun, "Traveling Neutral Disturbances," to be published in Geophys. Res. Lett. (1976).

S. H. Gross, "Galilean Satellites and Jovian Energetic Particles," Sci., 190, 564-566 (November 1975).

T. Hom, W. Kiszenick and B. Post, "Accurate Lattice Constants from Multiple Reflection Measurements. II. Lattice Constants of Germanium, Silicon and Diamond," J. Appl. Cryst., 8, 457 (August 1975).

P. Kersten and L. Kurz, "Bivariate m-Interval Classifiers with Application to Edge Detection," to be published in Information and Control.

I. Kupiec, S. Rosenbaum and L. B. Felsen, "Backscatter from Scattering Cells Located in a Random Medium," to be published in Radio Science.

L. Kurz and C. Mohwinkel, "Computer Picture Processing and Enhancement by Localized Operations," to be published in Computer Graphics and Image Processing.

P. E. Lagasse and A. A. Oliner, "Acoustic Flexural Mode on a Ridge of Semi-Infinite Height," Electron. Lett., 12, No. 1 (January 1976).

R. C. M. Li, A. A. Oliner and H. L. Bertoni, "Microwave Network Analyses of Surface Acoustic Waveguides, I: Flat Overlay Guides," to be published in IEEE Trans. Sonics and Ultrasonics.

B. S. Mangla and E. S. Cassedy, "Passive Mode-Locking of a Ruby Laser Using a Birefringent Crystal and a Fractional-Commensurate Cavity Length," Appl. Phys. (May 1976).

S. Markman, R. C. M. Li, A. A. Oliner and H. L. Bertoni, "Microwave Network Analyses of Surface Acoustic Waveguides, II: Rectangular Ridge Guides," to be published in IEEE Trans. Sonics and Ultrasonics.

A. A. Oliner, "Waveguides for Acoustic Surface Waves: A Review," Proc. IEEE, 64, No. 5, 615-627 (May 1976).

H. Schachter, K. S. Meng, W-C. Wang and F. A. Cassara, "An Application of Acoustic Surface Wave Devices to Communication Receivers," IEEE Trans. on Electron Devices (September 1976).

L. Shaw, "Busy Period Control of Queues Based on Waiting Times at Arrivals," to be published in Operations Research.

L. Shaw and M. L. Shooman, "Confidence Bounds and Propagation of Uncertainties in System Availability and Reliability Computations," to be published in Naval Research Logistics Quarterly.

L. Shaw and S. Sinkar, "Redundant Spares Allocation to Reduce Reliability Costs," Naval Research Logistics Quarterly, 23, No. 2, 179-194 (June 1976).

M. L. Shooman and M. I. Bolsky, "Types, Distribution and Test Correction Times for Programming Errors," IEEE Trans. on Reliability, R-25, No. 2, 69-70 (June 1976).

M. L. Shooman and A. K. Trivedi, "A Many-State Markov Model for Computer Software Performance Parameters," IEEE Trans. on Reliability, R-25, No. 2, 66-68 (June 1976).

T. Tamir, "Leaky Waves in Planar Optical Waveguides," Nouv. Rev. Optique, 6, 273-284 (September-October 1975).

W. C. Wang and K. S. Meng, "Surface Wave Correlation and Time Scaling in a Structure of BGO-Si-$LiNbO_3$," Appl. Phys. Lett., 27, No. 7, 375 (October 1975).

K. H. Yen and A. A. Oliner, "The Strip Acoustic Surface Waveguide: Comparison Between Measurement and Theory," Appl. Phys. Lett., 28, No. 7, 368-370 (April 1976).

D. C. Youla, J. J. Bongiorno, Jr. and H. Jabr, "Modern Wiener-Hopf Design of Optimal Controllers - Part I: The Single-Input-Output Case," IEEE Trans. on Automatic Control, AC-21, No. 1 (February 1976).

D. C. Youla, H. Jabr and J. J. Bongiorno, Jr., "Modern Wiener-Hopf Design of Optimal Controllers - Part II: The Multivariable Case," IEEE Trans. on Automatic Control, AC-21, No. 3 (June 1976).

CONTRIBUTIONS TO RECENT BOOKS AND TO PROCEEDINGS

L. Assadourian and L. M. Silber, "Crystalline Anisotropy of Cobalt Ferrite from Observations of Ferromagnetic Resonance," Proc. of 21st Conference on Magnetism and Magnetic Materials/AIP, Philadelphia, Pennsylvania (December 1975).

H. L. Bertoni and Y. L. Hou, "Effects of Damping in a Solid on Acoustic Beams Reflected at the Rayleigh Critical Angle," Proc. of 10th Nondestructive Testing Symposium, Southwest Research Institute, San Antonio, Texas, 136-142 (April 1975).

H. L. Bertoni, "A Survey of Optical Waveguides 1972-1974," a paper in Review of Radio Science, 1972-1974, S. A. Bowhill, Ed. (Brussels: International Union of Radio Science, 1975).

L. Birenbaum, I. T. Kaplan, W. Metlay, S. W. Rosenthal and M. M. Zaret, "Effects of 35 and 107 GHz Microwaves on the Rabbit Eye," Proc. of 1975 URSI Symposium, Boulder, Colorado (October 1975).

B. R. Cheo, "Truncation of Sheath Spiral Antennas," Proc. of 1975 URSI Symposium, Boulder, Colorado (October 1975).

S. Choudhary and L. B. Felsen, "Propagation in Inhomogeneous Slab Waveguides," Proc. of Microwave Theory and Techniques International Symposium, Cherry Hill, New Jersey (June 1976).

L. B. Felsen, "Beams and Modes in Optical Fibers," Proc. of 1975 URSI Symposium, Boulder, Colorado (October 1975).

L. B. Felsen, "Evanescent Waves," Proc. of Microwave Theory and Techniques International Symposium, Cherry Hill, New Jersey (June 1976).

L. B. Felsen, "Asymptotic Methods in Propagation and Scattering," Proc. of National Conference on Electromagnetic Scattering, Chicago, Illinois (June 1976).

L. B. Felsen, "Ray Methods for Evanescent Fields," Proc. of 6th European Microwave Conference, Rome, Italy (September 1976).

L. B. Felsen, "Transient Electromagnetic Wave Propagation and Diffraction in Non-dispersive and Dispersive Media," a chapter in Transient Electromagnetic Fields, Topics in Appl. Phys. Series, Vol. 10, L. B. Felsen, Ed. (New York: Springer-Verlag, 1976).

R. Foye and S. Barone, "Classical Theory of a Neutral Magnetic Dipole in a Static Electric Field," Proc. of Meeting of the American Physical Society, New York, New York, 21, 24 (Bull. Am. Phys. Soc., 1976).

S. H. Gross and G. V. Ramanathan, "Time Evolution of Primitive Hydrogen Atmospheres," Proc. of XVIth IUGG General Assembly, Grenoble, France (September 1975).

S. H. Gross, "The Onset of Instability Against Escape of an Atmosphere," Proc. of XVIth IUGG General Assembly, Grenoble, France (September 1975).

S. H. Gross, "Io's Ionosphere and the Sodium Clouds," Proc. of 1976 Meeting of the Division of Planetary Sciences/AAS, Austin, Texas (April 1976).

F. Haisch and R. Boorstyn, "Adaptive Message Switching Networks," Proc. of International Conference on Communications, Philadelphia, Pennsylvania, 38.7-38.13 (June 1976).

A. Kershenbaum and R. Boorstyn, "Centralized Teleprocessing Network Design," Proc. of IEEE National Telecommunications Conference, New Orleans, Louisiana (December 1975).

L. Kurz and G. Pasternack, "Non-parametric Classification Using Quadratic Quantile Statistics," Proc. of 1976 International Symposium on Information Theory, Ronneby, Sweden (June 1976).

E. Levi, "Linear Propulsion," a paper in Standard Handbook for Electrical Engineers, 11th Edition (New York: McGraw Hill, 1976).

N. Marcuvitz, "Quasiparticle Approach to Nonlinear and Turbulent Wave Problems," Proc. of International Electromagnetics Symposium, University of Illinois at Chicago Circle (June 1976).

N. Marcuvitz, "Characteristic Green's Function (and Eigenvectors) for Coupled Electromagnetic and Plasma Fields," Symposia Matematica, Vol. XVIII, 85-106 (Academic Press, 1976).

F. P. Mattar and M. C. Newstein, "Transverse Effects in the Propagation of Coherent Pulses," Proc. of 10th Congress of International Commission for Optics, Prague, Czechoslovakia (August 1975).

D. J. Mayer and L. Shaw, "The Stability of Occupancy Based Library-Type Traffic Control Systems," Digest of Papers presented at the 55th Transportation Res. Bd. Meeting, Washington, DC (January 1976).

P. Melman and S. Barone, "Asymptotic Solution of the Dirac Equation and the Inertial Mass of an Electron," Proc. of Meeting of the American Physical Society, New York, New York, 21, 24 (Bull. Am. Phys. Soc., 1976).

S. N. Mohanty and M. Adamowicz, "Proposed Measures for Evaluation of Software," Proc. of Symposium on Computer Software Engineering, MRI Symp. Series, Vol. XXIV, J. Fox, Ed. (New York: Polytechnic Press, 1976).

M. C. Newstein and F. P. Mattar, "Self-Focusing of Coherent Optical Pulses in Resonant Absorbers," Digest of Papers presented at the Meeting of the Optical Society of America, Boston, Massachusetts (October 1975).

A. A. Oliner and S. T. Peng, "Optical Slot and Strip Waveguides Formed by Metal Platings," Tech. Digest, Topical Meeting on Integrated Optics, Optical Society of America, Salt Lake City, Utah, 75 CH 1039-7QEC (January 1976).

A. A. Oliner and S. T. Peng, "Waveguides for Integrated Optics Formed by Metal Platings," Proc. of 6th European Microwave Conference, Rome, Italy (September 1976).

G. Osche and S. Barone, "On the Physical Interpretation of the Dirac-Pauli Theory of a Neutral Dipole in a Static Electric Field," Proc. of Meeting of the American Physical Society, New York, New York, 21, 24 (Bull. Am. Phys. Soc., 1976).

S. T. Peng, T. Tamir, D. W. Fradin and P. K. Cheo, "Effects of Imperfections on the Efficiency of Optical Couplers," Tech. Digest, Topical Meeting on Integrated Optics, Optical Society of America, Salt Lake City, Utah, WD-1-WD-4 (January 1976).

A. A. Sarris, M. Mor and E. J. Smith, "On Test Generation for Asynchronous Sequential Circuits," Proc. of 1976 International Symposium on Fault-Tolerant Computing, Pittsburgh, Pennsylvania (June 1976).

H. Schachter, J. Franck and D. Schilling, "Three Level Delta Modulation," Proc. of IEEE National Telecommunications Conference, New Orleans, Louisiana (December 1975).

H. Schachter, J. S. Sundresh and F. A. Cassara, "A Novel FM Detector for Suppression of Interchannel Interference," Proc. of 6th European Microwave Conference, Rome, Italy (September 1976).

L. Shaw and N. Wattanapanom, "Optimal Testing Times for Failure Detection when Tests Hasten Failure," Digest of Papers presented at the Operations Research Society Meeting, Philadelphia, Pennsylvania (March 1976).

M. L. Shooman and S. Natarajan, "Effect of Manpower Deployment and Error Generation on Software Reliability," Proc. of Symposium on Computer Software Engineering, MRI Symp. Series, Vol. XXIV, J. Fox, Ed. (New York: Polytechnic Press, 1976).

M. L. Shooman, M. Horodniceanu and E. J. Cantilli, "System Safety Applied to Transportation Systems," Proc. of Inter Society Conference on Transportation, Los Angeles, California (July 1976).

M. L. Shooman, "Recent Developments in Software Reliability - The State of the Art," Proc. of 13th IEEE Computer Society International Conference, Washington, DC (September 1976).

T. Tamir, Chapter 1, "Introduction," and Chapter 3, "Beam and Waveguide Couplers," Integrated Optics, T. Tamir, Ed. (Springer-Verlag, 1975).

T. Tamir, S. T. Peng and H. L. Bertoni, "Beam and Waveguide Couplers," Proc. of Electro-Optics International Laser Exposition, Anaheim, California (November 1975).

PATENTS

E. Levi, Magnetic Separation of Isotopes, U.S. Patent No. 3973926, August 1976.

RECENT BOOKS

"Proceedings of Symposium on Computer Software Engineering," General Chairman: A. A. Oliner, MRI Symposia Series, Vol. XXIV, J. Fox, Ed. (New York: Polytechnic Press, 1976).

TECHNICAL REPORTS

E. Banks, "Compounds with Defect Lattice Structure" (DAHC04-75-C-0096 for U.S. Army Research Office).

A. Benler, "Simulation and Estimation of a Stochastic Model for the Growth of Ridership of a new Transportation System" (POLY EE/EP76-008).

C. L. Hsu and L. Shaw, "Downtime Distributions Based on a Multivariate Exponential Distribution" (POLY EE/EP76-002).

E. Kunhardt and B. R. Cheo, "An Experimental and Theoretical Study of the Propagation of High Amplitude Pulses in a Bounded Magneto-Plasma" (POLY EE/EP76-011).

E. J. Lancevich and E. J. Smith, "The Systematic Design of Communications Line-Control Procedures" (POLY EE/EP76-010).

T. Lee and E. J. Smith, "On the Partitioning Representation of Finite Sets and its Application" (POLY EE/EP76-009).

E. Levi, "Preliminary Design Studies on Iron-cored Synchronously Operating Linear Motors" (POLY EE/EP76-005).

E. Levi, "A Preliminary Evaluation of Electrical Propulsion by Means of Iron-Cored Synchronously Operating Linear Motors" (POLY EE/EP75-011).

C. Marshall, "Contributions to the Theory of Component Availability" (POLY EE/EP76-004).

E. A. Mishkin, "Optically Generated Shocks" (POLY EE/EP75-009).

E. A. Mishkin, "Spectral Radiation from a Sphere" (POLY EE/EP76-006).

S. Natarajan and M. L. Shooman, "Effects of Manpower Deployment and Error Generation on Software Reliability" (POLY EE/EP76-007).

L. Shaw and M. L. Shooman, "Confidence Bounds and Propagation of Uncertainties in System Availability and Reliability Computations" (NOOO14-67-A-0438-0013; POLY EE/EP75-002).

M. L. Shooman and H. Ruston, "Summary of Technical Progress - Software Modeling Studies" (POLY EE/EP76-013).

P. C. Wang, W. Wang and R. F. Drenick, "Case Study of Critical Excitation and Response of Structures" (POLY EE/EP75-010).

N. Wattanapanom and L. Shaw, "Optimal Inspection Schedules for Failure Detection When Tests Hasten Failures" (POLY EE/EP76-001).

blank

DISTRIBUTION LIST*
Contract No. F44620-74-C-0056

Report on the Joint Services Electronics Program for the Period Ending 14 September 1976

DEPARTMENT OF DEFENSE

Defense Documentation Center
ATTN: DDC-TCA (Mrs. V. Caponio) (12)
Cameron Station
Alexandria, Virginia 22314

Asst. Dir, Electronics & Computer Sciences
Office of Director of
Defense Research and Engineering
The Pentagon
Washington, DC 20315

Office of the Director of Defense
Research and Engineering
Information Office Library Branch
The Pentagon
Washington, DC 20301

ODDR&E Advisory Group on Electron Devices
201 Varick Street
New York, New York 10014

Chief, R&D Division (340)
Defense Communications Agency
Washington, DC 20301

Director, National Security Agency
Fort George G. Meade, Maryland 20755
ATTN: Dr. T. J. Beahn

Institute for Defense Analysis
Science and Technology Division
400 Army-Navy Drive
Arlington, Virginia 22202

Dr. Stickley
Defense Advanced Research Projects Agency
ATTN: Technical Library
1400 Wilson Boulevard
Arlington, Virginia 22209

Dr. R. Reynolds
Defense Advanced Research Projects Agency
ATTN: Technical Library
1400 Wilson Boulevard
Arlington, Virginia 22209

DEPARTMENT OF THE AIR FORCE

AF/RDPS
The Pentagon
Washington, DC 20330

AFSC (LJ/Mr. Irving R. Mirman)
Andrews AFB
Washington, DC 20334

Directorate of Electronics and Weapons
HQ AFSC/DLC
Andrews AFB, Maryland 20334

Directorate of Science
HQ AFSC/DLS
Andrews AFB
Washington, DC 20334

LTC J. W. Gregory (5)
AF Member, TAC
Air Force Office of Scientific Research
(AFSC) AFSOR/NE
Bolling Air Force Base, DC 20332

Mr. Carl Sletten
RADC/ETE
Hanscom AFB, MA 01731

Dr. Richard Picard
RADC/ETSL
Hanscom AFB, MA 01731

Mr. Robert Barrett
RADC/ETS
Hanscom AFB, MA 01731

Dr. John N. Howard
AFGL/CA
Hanscom AFB, MA 01731

Dr. Richard B. Mack
RADC/ETER
Hanscom AFB, MA 01731

Documents Library (TILD)
Rome Air Development Center
Griffiss AFB, New York 13441

Mr. H. E. Webb, Jr (ISCP)
Rome Air Development Center
Griffiss AFB, New York 13441

Mr. Murray Kesselman (ISCA)
Rome Air Development Center
Griffiss AFB, New York 13441

Mr. W. Edwards
AFAL/TE
Wright-Patterson AFB, Ohio 45433

Mr. R. D. Larson
AFAL/DHR
Wright-Patterson AFB, Ohio 45433

Chief Scientist
AFAL/CA
Wright-Patterson AFB, Ohio 45433

HQ ESD (DRI/Stop 22)
Hanscom AFB, MA 01731

Professor R. E. Fontana
Head Dept. of Electrical Eng.
AFIT/ENE
Wright-Patterson AFB, Ohio 45433

Mr. John Mottsmith (MCIT)
HQ ESD (AFSC)
Hanscom AFB, MA 01731

LTC Richard J. Gowen
Professor
Dept. of Electrical Engineering
USAF Academy, Colorado 80840

AUL/LSE-9663
Maxwell AFB, Alabama 36112

AFETR Technical Library
P. O. Box 4600, MU 5650
Patrick AFB, Florida 32542

ADTC (DLOSL)
Eglin AFB, Florida 32542

HQ AMD (ROR/Col. Gooden)
Brooks AFB, Texas 78235

USAF European Office of Aerospace Research
Technical Information Office
Box 14, FPO, New York 09510

Dr. Carl E. Baum
AFWL (ES)
Kirtland AFB, New Mexico 87117

ASAFSAM/RAL
Brooks AFB, Texas

DEPARTMENT OF THE ARMY

HQDA (DAMA-ARZ-A)
Washington, DC 20310

Commander
US Army Security Agency
ATTN: IARD-T
Arlington Hall Station
Arlington, VA 22212

Commander
US Army Materiel Development and Readiness Command
ATTN: Technical Library Rm 7S 35
5001 Eisenhower Avenue
Alexandria, VA 22333

Commander
US Army Ballistics Research Laboratory
ATTN: DRXRD-BAD
Aberdeen Proving Ground
Aberdeen, MD 21005

Commander
Picatinny Arsenal
ATTN: SMUPA-TS-T-S
Dover, NJ 07801

US Army Research Office
ATTN: Dr. Hermann Robl
P. O. Box 12211
Research Triangle Park, NC 27709

US Army Research Office
ATTN: Mr. Richard O. Ulsh
P. O. Box 12211
Research Triangle Park, NC 27709

US Army Research Office
ATTN: Dr. Jimmie R. Suttle
P. O. Box 12211
Research Triangle Park, NJ 27709

US Army Research Office
ATTN: Dr. Horst Wittmann
P. O. Box 12211
Research Triangle Park, NC 27709

Commander/Director
Athospheric Sciences Laboratory (ECOM)
ATTN: DRSEL-BL-DD
White Sands Missile Range, NM 88002

Director
Electronic Warfare Laboratory (ECOM)
ATTN: DRSEL-WL-MY
White Sands Missile Range, NM 88002

Commander
US Army Armament Command
ATTN: DRSAR-RD
Rock Island, IL 61201

Director, Division of Neuropsychiatry
Walter Reed Army Institute of Research
Washington, DC 20012

Commander
USASATCOM
Fort Monmouth, NJ 07703

Commander
US Army R&D Group (Far East)
APO San Francisco, CA 96343

Commander
US Army Communications Command
ATTN: Director, Advanced Concepts Office
Fort Huachuca, AZ 85613

Commander
Frankford Arsenal
ATTN: Mr. George C. White, Jr.
Deputy Director, Pitman-Dunn Laboratory
Philadelphia, PA 19137

Commander
US Army Missile Command
ATTN: Chief, Document Section
Redstone Arsenal, AL 35809

Commander
US Army Missile Command
ATTN: DRSMI-RR
Redstone Arsenal, AL 35809

Commander
US Army Materials and Mechanics Research Center
ATTN: Chief, Materials Sciences Division
Watertown, MA 02172

Commander
Harry Diamond Laboratories
ATTN: Mr. John E. Rosenberg
2800 Powder Mill Road
Adelphi, MD 20783

Commandant
US Army Air Defense School
ATTN: ATSAD-T-CSM
Fort Bliss, TX 79916

Commandant
US Army Command and General Staff College
ATTN: Acquisitions, Lib Div
Fort Leavenworth, Kansas 66027

Dr. Hans K. Ziegler
Army Member, TAC/JSEP
US Army Electronics Command
(DRSEL-TL-D)
Fort Monmouth, NJ 07703

Mr. J. E. Teti (3)
Executive Secretary, TAC/JSEP
US Army Electronics Command
(DRSEL-TL-DT)
Fort Monmouth, NJ 07703

Director
Night Vision Laboratory, ECOM
ATTN: DRSEL-NV-D
Fort Belvoir, VA 22060

Project Manager
ARTADS
EAI Building
West Long Branch, NJ 07764

Commander
US Army White Sands Missile Range
ATTN: STEWS-ID-R
White Sands Missile Range, NM 88002

Director, TRI-TAC
ATTN: TT-AD (Mrs. Brieller)
Fort Monmouth, NJ 07703

Commander
US Army Communication Command
ATTN: CC-OPS-PD
Fort Huachuca, AZ 85613

COL Robert Noce
Senior Standardization Representative
US Army Standardization Group, Canada
Canadian Force Headquarters
Ottawa, Ontario, Canada K1A 0K2

Commander
US Army Electronics Command
ATTN: DRSEL-RD-O (Dr. W. S. McAfee) (13)
CT-L (Dr. R. Buser)
ML-O (Dr. H. S. Bennett)
ML-T (Mr. R. Kulinyi)
TL-B
VL-D
WL-D
TL-AM (Mr. N. Lipetz)
ML-M (Dr. F. Schwering)
TL-E (Dr. S. Kronenberg)
TL-E (Dr. J. Kohn)
TL-I (Dr. C. Thornton)
ML-B (Dr. S. Amoroso)
Fort Monmouth, NJ 07703

1 COPY TO EACH SYMBOL LISTED INDIVIDUALLY ADDRESSED

Project Manager
Ballistic Missile Defense Program Office
ATTN: DACS-BMP (Mr. A. Gold)
1300 Wilson Blvd
Washington, DC 22209

Dr. Sidney Ross
Technical Director
SARFA-TD
Frankford Arsenal
Philadelphia, PA 19137

DEPARTMENT OF THE NAVY

Office of Naval Research
Electronic and Solid State Sciences Program (Code 427)
800 N. Quincy
Arlington, Virginia 22217

Office of Naval Research
Code 200
Assistant Chief for Technology
800 N. Quincy
Arlington, Virginia 22217

Office of Naval Research
Information Sciences Program (Code 437)
800 N. Quincy
Arlington, Virginia 22217

Naval Research Laboratory
4555 Overlook Avenue, S. W.
Washington, DC 20375
ATTN: Codes 2627 (15)
4000
4105
5000
5200
5203
5210
5270
5300
5400
5460
5460
5500
5510
6400

Director
Office of Naval Research Branch Office
536 South Clark Street
Chicago, Illinois 60605

San Francisco Area Office
Office of Naval Research
760 Market Street, Room 447
San Francisco, California 94102

Dr. A. Laufer
Chief Scientist
Office of Naval Research Branch Office
1030 East Green Street
Pasadena, California 91101

NOTE: 1 COPY TO EACH ADDRESSEE UNLESS OTHERWISE INDICATED

*The Joint Services Technical Advisory Committee has established this list for the regular distribution of reports on the electronics research program of the Polytechnic Institute of New York. Additional addressees may be included on their written request to:

Mr. John E. Teti, (DRSEL-TL-DT)
Executive Secretary, TAC/JSEP
U.S. Army Electronics Command
Fort Monmouth, New Jersey 07703

An appropriate endorsement by a Department of Defense sponsor is required except on a request from a Federal Agency.

Director
Office of Naval Research Branch Office
715 Broadway, 5th Floor
New York, New York 10003

Mr. L. W. Sumney
Naval Electronics Systems Command
NC #1
2511 Jefferson Davis Highway
Arlington, Virginia 20360

Mr. R. Fratila
Naval Electronics Systems Command
NC #1
2511 Jefferson Davis Highway
Arlington, Virginia 20360

Mr. H. Butler
Naval Electronics Systems Command
NC #1
2511 Jefferson Davis Highway
Arlington, Virginia 20360

Dr. H. J. Mueller
Naval Air Systems Command
JP #1
1411 Jefferson Davis Highway
Arlington, Virginia 20360

Capt. R. B. Meeks
Naval Sea Systems Command
NC #3
2531 Jefferson Davis Highway
Arlington, Virginia 20362

Commander
Naval Surface Weapons Center
ATTN: Technical Library
Silver Spring, Maryland 29010

Naval Surface Weapons Center
ATTN: Code 212
Silver Spring, Maryland 29010

Officer-in-Charge
Naval Surface Weapons Center
Dahlgren Laboratory
Dahlgren, Virginia 22448

Naval Air Development Center
ATTN: Technical Library
Johnsville
Warminster, Penn. 18974

Commander
Naval Avionics Facility
Indianapolis, Indiana 46241
ATTN: D/035 Technical Library

Naval Missile Center
Technical Library
Code 5632.2
Point Mugu, California 93042

Naval Weapons Center
ATTN: Technical Library, Code 533
China Lake, California 93555

Naval Weapons Center
ATTN: Code 6010
China Lake, California 93555

Naval Weapons Center
Mathematics Division
China Lake, California 93555

Naval Training Equipment Center
Technical Library
Orlando, Florida 32813

Naval Research Laboratory
Underwater Sound Reference Division
Technical Library
P. O. Box 8337
Orlando, Florida 32806

Naval Underwater Sound Center
Technical Library
New London, Conn. 06320

U. S. Naval Oceanographic Office
Library - Code 1600
Washington, DC 20373

Commandant, Marine Corps
Scientific Advisor (Code AX)
Washington, DC 20380

Dr. Gernot M. R. Winkler
Director, Time Service
US Naval Observatory
Mass. Ave. at 34th St. N.W.
Washington, DC 20390

Naval Postgraduate School
Technical Library
Monterey, California 93940

Naval Electronics Laboratory Center
Technical Library
San Diego, California 92152

Naval Electronics Laboratory Center
ATTN: Code 2600
San Diego, California 92152

Naval Electronics Laboratory Center (10)
San Diego, California 92152
ATTN: Codes 2000
2200
2300
3000
3200
3300
4800
5000
5200
5300

Naval Undersea Center
Technical Library
San Diego, California 92152

Naval Ship Research and Development Center
David W. Taylor
Code 522.1
Bethesda, Maryland 20084

Office of Chief of Naval Operations
OP-941D3/M15 Planning Branch
NDP-9160, Pentagon
Washington, DC 20350

OTHER GOVERNMENT AGENCIES

Mr. F. C. Schwenk, RD-T
National Aeronautics & Space Administration
Washington, DC 20546

Los Alamos Scientific Laboratory
ATTN: Reports Library
P. O. Box 1663
Los Alamos, New Mexico 87544

M. Zane Thornton
Deputy Director Institute for
Computer Sciences & Technology
National Bureau of Standards
Washington, DC 20234

Director, Office of Postal Technology (R&D)
US Postal Service
11711 Parklawn Drive
Rockville, Maryland 20852

NASA Lewis Research Center
ATTN: Library
21000 Brookpark Road
Cleveland, Ohio 44135

Library -R51
Bureau of Standards
Acquisition
Boulder, Colorado 80302

MIT Lincoln Laboratory
ATTN: Library A-082
P. O. Box 73
Lexington, Massachusetts 02173

Dr. Jay Harris
Program Director, Devices and Waves Program
NSF
1800 G Street
Washington, DC 20550

Dr. Howard W. Etzel
Deputy Director, Div of Materials Rsch
NSF
1800 G Street
Washington, DC 20550

Dr. Dean Mitchell
Program Director, Solid-State Physics
Division of Materials Research
National Science Foundation
1800 G Street
Washington, DC 20550

NON-GOVERNMENT AGENCIES

Director
Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

Director
Microwave Research Institute
Polytechnic Institute of New York
Long Island Graduate Center, Route 110
Farmingdale, New York 11735

Abbt, Dir, Microwave Research Institute
Polytechnic Institute of New York
333 Jay Street
Brooklyn, New York 11201

Director
Columbia Radiation Laboratory
Dept of Physics
Columbia University
538 West 120th Street
New York, New York 10027

Director
Coordinated Science Laboratory
University of Illinois
Urbana, Illinois 61801

Director
Stanford Electronics Laboratory
Stanford University
Stanford, California 94305

Director
Microwave Laboratory
Stanford University
Stanford, California 94305

Director
Electronics Research Laboratory
University of California
Berkeley, California 94720

Director
Electronics Sciences Laboratory
University of Southern California
Los Angeles, California 90007

Director
Electronics Research Center
The University of Texas at Austin
Engineering-Science Bldg 112
Austin, Texas 78712

Director of Laboratories
Division of Engineering & Applied Physics
Harvard University
Pierce Hall
Cambridge, Massachusetts 02138

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) PROGRESS REPORT No. 41 TO THE JOINT SERVICES TECHNICAL ADVISORY COMMITTEE		5. TYPE OF REPORT & PERIOD COVERED Scientific Interim 6. PERFORMING ORG. REPORT NUMBER POLY-MRI-452.41-76
7. AUTHOR(s) Arthur A. Oliner		8. CONTRACT OR GRANT NUMBER(s) F 44620-74-C-0056
9. PERFORMING ORGANIZATION NAME AND ADDRESS Polytechnic Institute of New York Microwave Research Institute 333 Jay Street, Brooklyn, NY 11201		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Project No. 4751 61102F 681305
11. CONTROLLING OFFICE NAME AND ADDRESS Joint Services Electronics Program through Dir. of Electronic and Solid State Sciences Air Force Office of Scientific Research (NE) Bolling AFB, Washington, DC 20332		12. REPORT DATE November 1976 13. NUMBER OF PAGES 639
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

1. This document has been approved for public release and sale; its distribution is unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

TECH, OTHER

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Acoustic Correlators
Antennas
Communications
Computer Communications
Computers
Control Theory
Data Processing
Electric Power Engineering
Electromagnetics
Electronics
Electrophysics
Energy Conversion

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This report is the forty-first in a series of progress reports to the Joint Services Technical Advisory Committee since inception of the Joint Services Electronics Program at the Polytechnic Institute of Brooklyn in July 1955. The report, now being issued annually, summarizes research accomplished under the aegis of the Microwave Research Institute and reflects the impact of the Joint Services Electronics Program on the research activities of faculty and students of the Institute. The program covers a broad spectrum ranging from basic theoretical physics, mathematics, and engineering, to experimental

over

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

19.

Micro-Electronics
Microwave Acoustics
Network Theory
Optics
Quantum Electronics
Software Reliability and Engineering
Solid State and Materials
Systems Control
Waveguide Techniques
Wave-Matter Interactions

20.

investigations involving basic measurements, development of devices, and materials.

Each activity reports in summary fashion on specific results obtained during the report period, 15 September 1975 through 14 September 1976, with individual acknowledgement of the sponsorship which has contributed to the reported work. The report is organized into two major divisions. The first, Electrophysics, includes the topics of: Electromagnetics; Acoustics; Optics; Quantum Electronics; Solid State and Materials; Wave-Matter Interactions; and Electric Power Engineering. The second, Systems, includes the topics of: Communications; Computer and Computer-Communications Networks; Safety, Reliability and Software Engineering; Systems, Control and Networks; and Data Processing.

Topics representative of most of the Microwave Research Institute programs have been included in this report. The various sponsoring agencies and their contract numbers are individually acknowledged at the end of each item.

This comprehensive annual report to the Joint Services Technical Advisory Committee is made possible through the support of the Air Force Office of Scientific Research of the Air Force Systems Command; the Department of the Army, Army Research Office; and the Department of the Navy, Office of Naval Research; under Contract F44620-74-C-0056, Project No. 4751.

Reproduction in whole or in part is permitted for any purpose of the United States Government.

Qualified requestors may obtain additional copies from the Defense Documentation Center; all others should apply to the Clearinghouse for Federal Scientific and Technical Information.